Manuscript papers prepared for the workshop

"Innovation on new digital exponential technologies towards the generation of Business Models"

Alicante, Archivo Histórico Provincial, 2-3rd September, 20221

Co-funded by the
Creative Europe Programme
of the European Union

Table of contents

Introduction

Innovation on new digital exponential technologies in the archives

Tatjana Hölzl, Communication Management / Financial Administration. ICARUS - International Centre for Archival Research

Guillermo José Morán Dauchez Deputy Director of the General Archive of the Indies

The workshop “Innovation on new digital exponential technologies in the archives” created by the Spanish State Archives within the framework of the European Digital Treasures project was dedicated to the keywords innovation & technology in archives.

It was held as an hybrid event on 2nd & 3rd of September at the Provincial Historical Archive of Alicante (Spain) and could be followed remotely via YouTube and on the website of the European Digital Treasures project.

With 12 speakers from 5 different EU countries, each a professional on their field, the workshop combined inspiring topics like new methods of digital imaging, automatic processing of historical data, building historical big data from archival records. The workshop informed about cross-discipline research projects, language technology, geospatial data & much more.

These speakers have presented ongoing and groundbreaking projects in which a very miscellaneous array of professional backgrounds and previous experiences has been put in play, to offer what innovation really is: new combinations of ideas and perspectives, both new and already established ones, that lead to new functional solutions for the above said challenges: the application of machine learning to very time consuming archival tasks such as indexing, or the use of high end medical technology in the access without damage records in dramatic conditions of preservation; or the automatic rearrangement of data in order to satisfy the needs of certain population groups have been some of the developments presented.

Questions like “What do medicine & digital humanities have in common?” arise. Scientific disciplines, archives & technology cannot be divided anymore – but why? What is deep learning and how can historical documents be preserved with noninvasive imaging methods? The workshop gives an answer to all those and many more questions.

Keeping records of transactions, historically, appears to have been the first application of writing -that amazing piece of technology, so powerful that it was then even deemed as magical, and is so common, yet so transcendent, nowadays .

Therefore, we must conclude that archives, at least as an implicit necessity, are as ancient as scripture itself: for a record to be kept, not only it must be materialized through writing, but

also preserved and conveniently sorted, so it could be retrieved whenever needed. Why would it be written in the first place, if the latter couldn't be satisfied?

Preservation and retrieval are the two main vectors of archivistics as a discipline; and until very recently, they have been arguably antagonistic concepts: by the mere fact of retrieving records to be consulted, a harm -as small as it can be-, is being made in their materiality. A minimal harm that will add up with every iteration until the loss of a record is complete. In the other hand, why would records be kept -or even created at all- if it is not to be retrieved and consulted?

Digital imaging and computing have, for sure, established a bridge, a compromise, between these two forces -use and preservation- in the last two or three decades. We can even assert that it has magnified the possibilities of these two realms, offering solutions that were self-evident up to a certain point, but also presenting new perspectives, the far sight of new possible solutions, in a world -in a society-, were the demand of information – public, reliable and transparent information- is in permanent and exponential growth, and is confronting archives and archivists with brand new challenges.

It is in the very hope of keeping up with, and even leading, society demands and expectations, that a whole profession is investing huge efforts in adopting an innovative mindset of which this workshop has given us one of its more brilliant examples.

Aggregation as a service

Automatic topic detection and collaborative topic tagging in Archives Portal Europe's multilingual environment

Kerstin Arnold

Abstract. *Archives Portal Europe (www.archivesportaleurope.net) is a comprehensive and open resource on archives from and about Europe, that currently holds archival descriptions from more than 30 countries and in more than 20 languages. Following traditional approaches of archival description, the portal allows users to access the documents via the contextual entities of the records creators and the holding repositories, next to a general keyword search. To evaluate options for subject- or topic-based access points, Archives Portal Europe is working on an automated cross-lingual topic detection tool that aims at enabling users to identify relevant documents related to a topic well beyond the narrowness of direct keyword matching. Synergising different approaches for concept-based and entity-based topics, the tool then also is meant to allow for active topic tagging in order to improve coverage of topic-based relations between the heterogeneous and multilingual documents present in Archives Portal Europe. Building on the current status quo in the portal, this paper presents the tool's set-up, initial results from the proof-of-concept phase, and next steps envisaged during alpha and beta development of the tool, which will be made available as Open Source to also be of benefit for other, similar initiatives in the cultural heritage sector.*

1 Background

In the mid-2000s, the European Union saw its single largest enlargement, when ten countries joined the bloc on 1 May 2004, with two more being added on 1 January 2007. This change had its effects in all areas of life, including the archives domain. On 6 May 2003, the Council of the European Union had issued its resolution on archives in the Member States (OJ 2003/C113/2) (European Union 2003), resulting in the establishment of the European Archives Group (EAG). During the following two years, the EAG, in collaboration with the European Board of National Archivists (EBNA), worked on what should become the Report on Archives in the enlarged European Union (European Union 2005), adopted by the Council in its recommendation on priority actions to increase cooperation in the field of archives in Europe (OJ 2005/L312/55) from 14 November 2005. The report included recommendations for actions in different areas, with

> "part two [...] deal[ing] with institutional, technical and professional aspects of access to archives. Particular emphasis is placed on [...] finding aids and archival description; access on line and new research tools; setting up an Internet Gateway/Portal to documents and archives in Europe and cooperation with European networking projects in this field [...]."
> (European Union 2005)

Thus, the idea for Archives Portal Europe (www.archivesportaleurope.net) was born.

Funded by the European Commission in two rounds from 2009 to 2012 and again from 2012 to 2015, Archives Portal Europe is now managed and developed further by the Archives Portal Europe Foundation (APEF), who took over all responsibilities and rights from the APEx (Archives Portal Europe network of excellence, www.apex-project.eu) and APEnet (Archives Portal Europe network, www.apenet.eu) projects in October 2015. The foundation has its physical headquarters in The Hague, Netherlands, but works with a remote and distributed core team of staff. It is supported financially by national archives, national archives administrations, and national archives aggregators from 21 countries in the role of foundation associates.

Archives Portal Europe is a comprehensive and open resource on archives from and about Europe, enabling new knowledge and new connections to be made. Its network represents a community of like-minded archives and cultural heritage professionals dedicated to the importance of sharing the continent's shared history and heritage. At the time of writing, the portal aggregates descriptions of more than 600,000 archival collections, with the majority being described up to item level and including links to digital archival objects, where applicable. The content published on the portal comes from over 1,100 institutions located in more than 30 countries and is made available in more than 24 languages and currently five different alphabets.

2 Approaches to Archival Description

In the way in which Archives Portal Europe presents the content made available on its platform, it very much follows the traditional approaches of archival description: the agents involved with the archival documents throughout their lifecycle, the records creators, their activities and tasks, give a collection of archival documents their initial structure, classification, and grouping, which provides further insight in how documents relate to each other and hence is usually kept as is when transferring documents into the archives. The holding repository, on the other hand, will – while staying true to the original contextualisation – add further contextual information relevant to its function of preserving the archival records and making them accessible and available to the public.

The classic statement that "the archive arises as a consequence of the activities of the person who formed it [because] the documents can only be understood from the point of view of the task involved" (Muller et al. 1940/2003, xx-xxi) is furthermore extended by an hierarchical approach to archival description. Here, "the fonds forms the broadest level of description [and] the parts form subsequent levels, whose description is often only meaningful when seen in the context of the description of the whole of the fonds" (International Council on Archives 2000, 8). This adds to the complexity of navigating archival description that users are presented with, as the approach via a multilevel description also means that "information that is common to the component parts [is given at the highest appropriate level only and is] not repeat[ed] at a lower level of description" (International Council on Archives 2000, 12).

Especially this last aspect of archival description leads to the question, how archives can embark on presenting the material they hold and the descriptions thereof in conjunction with and in relation to items from other cultural heritage institutions such as libraries and museums, where subject- and object-based approaches to description have traditionally been more common. And how archives can meet the expectations of users, who – in today's digital world of information retrieval – have grown accustomed to searching by subject-based access points even more.

3 Traditional and New Access Points to Archival Materials

3.1 Access Points in Archival Description Standards

While ISAD(G), the General International Standard Archival Description developed and maintained by the International Council on Archives "to be used in conjunction with existing national standards or as the basis for the development of national standards" (International Council on Archives 2000, 7), acknowledges the importance of access points for information retrieval in general, its main focus is on access points related to the agents that are named as records creators. With regard to other access points, the standard refers to national and language-specific developments as well as to more general conventions and frameworks that "are useful when developing and maintaining controlled vocabularies: ISO 5963 [...], ISO 2788 [...] and ISO 999 [...]" (International Council on Archives 2000, 9).[1]

Records in Contexts – A Conceptual Model for Archival Description (RiC-CM), the emerging new model for describing archives which is meant "to reconcile, integrate, and build on the four existing standards" (International Council on Archives 2012-2021)[2] extends the traditional approach by also including entities such as Event and Place, but defaults back to the general base entity Thing when it comes to connecting the archival documents with "all possible concepts, material things, or events within the realm of shared human experience and discourse [...] that are of primary interest to records managers and archivists, as well as other entities used in the description of the primary entities" (International Council on Archives 2021, 19). The accompanying Records in Contexts - Ontology (RiC-O), on the other hand, provides a whole range of additional classes for indexing archival description such as Type, Language, Physical Location, and Coordinates to only name a few.

3.2 Access Points when Encoding Archival Descriptions

Following the existing international and/or national standards, conventions, and rules for archival description, most archival management systems will provide options to identify, name, and potentially describe agents who are of importance in the context of archival documents; to certain extent, they will also include fields to capture information about other access points, such as subjects, on all levels of description, though these will often be bound to local or national vocabularies only.

The addition of such access points as part of the archival descriptions will also depend on other aspects such as whether or not it is part of the archival description tradition of an institution or country and the question of resources available to create detail-level descriptions to start with. Only in (versions of) archival management systems that have been developed more recently there will be a functionality that allows for the inclusion of references to

---

1 For the ISO standards mentioned see: ISO 5963:1985, https://www.iso.org/standard/12158.html; ISO 2788:1986, revised by ISO 25964-1:2011, https://www.iso.org/standard/53657.html, and ISO 25964-2:2013, https://www.iso.org/standard/53658.html; ISO 999:1996, https://www.iso.org/standard/5446.html, all last accessed on 1 September 2021.

2 Next to ISAD(G), the four standards mentioned here include: ISAAR(CPF) https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd; ISDF, https://www.ica.org/en/isdf-international-standard-describing-functions; ISDIAH, https://www.ica.org/en/isdiah-international-standard-describing-institutions-archival-holdings, all last accessed on 1 September 2021.

international vocabularies such as the Library of Congress Subject Headings (LCSH), the Getty Art & Architecture Thesaurus (AAT), or the UNESCO Thesaurus.

3.3 Access Points when Aggregating Archival Descriptions

The diversity of approaches also means that subject-based access points in archival descriptions present a specific challenge for aggregation initiatives, especially for those gathering materials from more than one country such as Archives Portal Europe.

Therefore, in addition to displaying existing subject headings as part of the archival descriptions and indexing these for the general keyword search in the portal, Archives Portal Europe has created a central, overarching topic-based approach that allows archival institutions to either make use of the appropriate encoding in their metadata or to add the relation between their archival documents and a specific subject as part of the central data processing. By this, users are offered access to the archival material via a list of predefined subject terms that connect documents in a variety of languages based on the mechanisms set out in the back-end of Archives Portal Europe.

However, the current process is not very flexible with regards to extending the list of subject terms that are available, and it furthermore relies on a manual intervention by the contributing institutions themselves during data processing, which leads to a rather inconsistent representation and coverage of these central topics across all countries and thereby languages connected to Archives Portal Europe at the moment.

4 Automated Topic Detection in a Multilingual Environment

Given the status quo as described in the previous section, Archives Portal Europe, in collaboration with an external developer and data scientist and supported by King's College London, has initiated a Research & Development project back in 2020 to evaluate the possibilities of applying automatic topic detection to its multilingual environment. Following the promising results of the initial proof-of-concept phase (see chapter 4.5), this work is currently being extended in the context of Archives Portal Europe's contribution to the Europeana Digital Service Infrastructure.

4.1 Objectives

The aim is to exploit methods of Natural Language Processing (NLP) in a supervised approach in order to train a tool that will help users as well as contributors to Archives Portal Europe in identifying materials related to a topic of their interest. In the short-term, this will be based on the topics that already exist in the central system of Archives Portal Europe, but it is the intention to also enable the work on and creation of new topics in the medium-term with the help of automated topic detection.

In the long-term, it is envisaged to add a second step following on from the identification of relevant documents: the possibility for contributors to Archives Portal Europe as well as its users to flag up these documents for active tagging with an appropriate subject heading, either in the central system or – ideally – at the source of the data, using international Linked Open Data vocabularies as a basis.

While developed in the context of and with the data aggregated by Archives Portal Europe, the approaches followed and the functionalities provided by the final tool, which will be made available as Open Source, are meant to be applicable to any heterogeneous and multilingual dataset from the cultural heritage and related domains.

4.2 Research Background

The project builds on work conducted in the Digital Humanities during the past two decades with a focus on adopting NLP methods for identifying topics in a supervised approach. This technique has been found suitable based on the circumstance that Archives Portal Europe (1) already provides a predefined list of topics, (2) works with a set of materials that have been manually annotated with the relevant topics, and (3) comes with a relatively large amount of such pre-annotated materials, approximately 2 million documents in total, to train the tool on documents in different languages.

Especially the last point qualifies the Archives Portal Europe case for a supervised approach, for which having a large enough dataset for training purposes often is the biggest obstacle, while the method generally offers a reliable performance for topic detection tasks (see for instance the experiments conducted by Merz et al. 2016 and by Glavaš et al. 2017 on the Manifesto Corpus). The alternative, i.e. an unsupervised approach, has been ruled out on the basis of the first and second points mentioned above: in a case like the one of Archives Portal Europe, in which the user already knows the topics contained in the collection, it can prove difficult to employ an unsupervised approach (Owens 2012), especially when applying Latent Dirichlet Allocation topic models, the results of which are often extremely hard to interpret (Chang et al. 2009) and are not always straightforward to align with our common notion of topics.

4.3 Cross-lingual Topic Classification for Concepts

Special emphasis has been given to supervised set-ups using a Support Vector Machine (SVM) (Joachims 2002) to train an algorithm, where each document is represented as a single feature-vector capturing the "meaning" of its content. In the case of Archives Portal Europe, we consider the descriptive units[3], i.e. the constituent components of an archival finding aid, as such "documents", representing each of them as the averaged vector of all its words, and thereby obtaining a single "document embedding" for each description. In order to address the unbalanced representation of languages in the dataset (see section 3.3), we apply Fast-Text word-embeddings and align these in a common cross-lingual "semantic" space by the project MUSE (Conneau et al. 2017-2018) to better represent all languages present in our dataset[4]. This approach has achieved really high performance, identifying the correct topical label for the materials in over 90% of the cases.[5] We additionally have ensured that the classifier was correctly distinguishing between topics, and not languages, by conducting an in-depth error analysis.

---

[3] The tool uses the Solr results in JSON format for each of these "documents", where some major parts of the archival description are captured in singular fields (e.g., the title of the unit itself or of the upper hierarchical levels that this unit is a part of). However, other parts of the archival descriptions are only included in a placeholder field of the Solr index, capturing all additional metadata that might be part of the original EAD-XML file. This is currently not part of the "document" as used by the tool.

[4] The dataset used in the proof-of-concept phase included documents in Finnish, French, German, Latvian, and Polish. Furthermore, we added English and Italian as supported languages for user queries. In the alpha phase, April to August 2021, this was extended further to also include Hebrew, Latvian, Russian, Spanish, Swedish.

[5] We obtained over 0.9 of both micro and macro F1-score, which is the harmonic mean of precision and recall. To know more see the documentation of the metrics on Scikit-learn, the library we adopted: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html.

While the proof-of-concept tool only allowed for single term keywords, this has been extended during the alpha phase to enable searches with the Boolean operators AND, OR, and AND NOT, as well as the use of wildcards like the asterisk replacing one or more letters of a search term. In such search scenarios, the tool will also show which terms have been included based on regular expressions running in the background while the search is conducted. It should be noted that at the time of writing this article, this functionality had only been added as a new extension to the tool and hence will require some more detailed testing to confirm its impact.

4.4 Extension to Entity-based Topics

In addition to searching for concepts, the tool also offers the option to search for entities across languages. Instead of relying on cross-lingual embeddings, the retrieval function first maps the entity inserted by the user as a query to its equivalent in Wikidata (when present). Next, it retrieves all name variations in the other languages under study,[6] and finally searches for their occurrence in the corpus. In the most recent alpha phase, this has been extended to also connect to the Virtual International Authority File, VIAF, and to include additional name variations from there. While this function is an early prototype and does not fully rely upon entity disambiguation approaches yet, we consider it useful as an additional way of exploring the collection.

4.5 Initial Results from the Proof-of-concept Phase

Tests with the proof-of-concept version of the tool have used a set of 457,538 documents already tagged with 1 of 9 topics[7] selected based on the following criteria:

- Balancing topics that cover a language broadly enough to learn from with topics that include documents in more than one language in order to address the multilingual character of Archives Portal Europe;
- Having topics of varied size and scope;
- Including topics that are entity-based as well as topics that are concept-based to address the two main approaches in archival research: persons/places on the one hand, subjects/themes on the other.

With regard to both, the concept search and the entity search, the evaluation of the tool's predominant function at this initial stage, the discovery of documents relevant to a topic, has given promising results. The tool allows the user to identify relevant documents related to a topic well beyond the narrowness of direct keyword matching and it has shown good results as well from not very largely annotated topics, which opens the door for smaller scale projects in future.

Nevertheless, it should be noted that, with the test data set only representing a rather small sample of the repository of Archives Portal Europe (25% of all documents tagged with a topic, and 0.16% overall), our experiments are only a first attempt towards a very challenging goal, and we plan to work with an interaction of supervised and unsupervised methods in future experiments, in order to tackle these challenges in a more comprehensive way.

---

[6] We pre-process name variations leaving life dates aside for persons or other characteristics sometimes included in brackets.

[7] For the alpha phase the data set was extended by four additional topics, reaching a total of about 675,000 documents being included.

5 Next Steps

Apart from the confirmation of the proof-of-concept, the initial results also have allowed us to elaborate on areas of further investigation and future development. Based on this, we have enlarged the sample data both in terms of topics under consideration and of available languages for the alpha phase, have enabled Boolean operators and wildcards in the search functionalities of the tool, and have started with the integration of other vocabularies and ontologies next to Wikidata along with an initial option for entity disambiguation.

The tool has been redesigned as a web application[8] including the display of other data from the documents, e.g. dates and the country where the contributing institution is located, that might be useful in determining whether a search result is relevant to a specific topic of interest or not. The tool now also allows for an export of results in CSV format for further analysis offline.

One of the next steps will include more detailed testing of these latest extensions and new functionalities in the context of workshops to be held with members of the archives community as well as with researchers in late 2021 and in early 2022. These workshops might concentrate on a specific topic or a certain language or language family in order to extend their representation in the Archives Portal Europe's data set, or they might test the tools functionality more generally without predefining a topic or language context.

In terms of developments, the focus will be on the inclusion of further vocabularies and ontologies such as the LCSH, the Getty AAT, or GeoNames, and the extension to the full data set of Archives Portal Europe. Furthermore, the beta phase will look into making use of the tool's results from the aggregating perspective of Archives Portal Europe itself, i.e. with regard to enabling Linked (Open) Data connections based on the entities identified via the tool and with regard to making such enriched metadata available in some way to the portal's users and contributors, and from the perspective of a contributor to Archives Portal Europe, i.e. with regard to transforming the results brought back by the tool into actual topic taggings in order to increase the representation of subject- or topic-based relations in the source data.

References

[1] Chang, Jonathan, and Gerrish, Sean, and Wang, Chong, and Boyd-Graber, Jordan L., and Blei, David M.. 2009. "Reading Tea Leaves – How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems 22 (NIPS 2009).* http://papers.nips.cc/paper/3700-Reading-tea-leaves-how-humans-interpret-topic-models.pdf (last accessed on 1 September 2021).

[2] Conneau, Alexis, and Lample, Guillaume, and Ranzato, Marc'Aurelio, and Denoyer, Ludovic, and Jégou, Hervé. 2017-2018. *Word Translation Without Parallel Data.* http://arxiv.org/abs/1710.04087 (last accessed on 1 September 2021).

[3] Europeana. 2018-2021. *Europeana Digital Service Infrastructure.* https://pro.europeana.eu/project/europeana-dsi-4 (last accessed on 1 September 2021).

[8] The public version of the tool is available at http://topicdetection.archivesportaleurope.net/. This will not always show the latest status of the development, but represents the most current stable version. All developments can be followed on GitHub at https://github.com/ArchivesPortalEuropeFoundation/Topic-Detection/.

[4] European Union. 2003. *Council Resolution of 6 May 2003 on archives in the Member States.* https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2003:113:0002:0002:EN:PDF (last accessed on 1 September 2021).
[5] European Union. 2005. *Report on Archives in the enlarged European Union.* https://ec.europa.eu/transparency/regdoc/rep/1/2005/EN/1-2005-52-EN-F1-2.Pdf (last accessed on 1 September 2021).
[6]European Union. 2005. *Council recommendation on priority actions to increase cooperation in the field of archives in Europe.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32005H0835&from=EN (last accessed on 1 September 2021).
[7] GeoNames. https://www.geonames.org/. (last accessed on 1 September 2021).
[8] Getty Art & Architecture Thesaurus. https://www.getty.edu/research/tools/vocabularies/aat/index.html (last accessed on 1 September 2021).
[9]Glavaš, Goran, and Nanni, Federico, and Ponzetto, Simone Paolo. 2017. "Cross-Lingual Classification of Topics in Political Texts." *Proceedings of the Second Workshop on NLP and Computational Social Science*: 42–46. Vancouver: Association for Computational Linguistics.
[10] International Council on Archives. 2000. *ISAD(G) – General International Standard Archival Description. 2nd edn.* Ottawa. https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf (last accessed on 1 September 2021).
[11] International Council on Archives. 2004. *International Standard Archival Authority Records – Corporate Bodies, Persons, and Families (ISAAR(CPF)). 2nd edn.* Paris. https://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf (last accessed on 1 September 2021).
[12] International Council on Archives. 2007. *International Standard Description of Functions (ISDF). 1st edn.* Paris. https://www.ica.org/sites/default/files/CBPS_2007_Guidelines_ISDF_First-edition_EN.pdf (last accessed on 1 September 2021).
[13] International Council on Archives. 2008. *International Standard Description of Institutions with Archival Holdings (ISDIAH). 1st edn.* Paris. https://www.ica.org/sites/default/files/CBPS_2008_Guidelines_ISDIAH_First-edition_EN.pdf (last accessed on 1 September 2021).
[14] International Council on Archives. 2012-2021. *Records in Contexts – Conceptual Model (RiC-CM) Homepage*. https://www.ica.org/en/egad-ric-conceptual-model (last accessed on 1 September 2021).
[15] International Council on Archives. 2017-2021. *Records in Contexts – Ontology (RiC-O).* https://www.ica.org/standards/RiC/RiC-O_v0-2.html (last accessed on 1 September 2021).
[16] International Council on Archives. 2021. *Records in Contexts – Conceptual Model v0.2.* https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf (last accessed on 1 September 2021).
[17] ISO 999:1996, Information and documentation – Guidelines for the content, organization and presentation of indexes. https://www.iso.org/standard/5446.html (last accessed on 1 September 2021).
[18] ISO 5963:1985. Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms. https://www.iso.org/standard/12158.html (last accessed on 1 September 2021).

[19] ISO 25964-1:2011, Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval. https://www.iso.org/standard/53657.html (last accessed on 1 September 2021).
[20] ISO 25964-2:2013, Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies. https://www.iso.org/standard/53658.html (last accessed on 1 September 2021).
[21] Joachims, Thorsten. 2002. *Learning to Classify Text Using Support Vector Machines.* Boston: Springer.
[22] Library of Congress Subject Headings. https://id.loc.gov/authorities/subjects.html (last accessed on 1 September 2021).
[23] Merz, Nicolas, and Regel, Sven, and Lewandowski, Jirka. 2016. "The Manifesto Corpus – A new resource for research on political parties and quantitative text analysis." *SAGE Journals, Research & Politics*. https://journals.sagepub.com/doi/10.1177/2053168016643346 (last accessed on 1 September 2021).
[24] Muller, Samuel, et al. 1940. *Manual for the arrangement and description of archives. 2nd edn. Translated by Arthur H. Leavitt.* New York: The H.W. Wilson Company. - Reissued 2003. Chicago: Society of American Archivists.
[25] Owens, Trevor. 2012. *Discovery and Justification are Different – Notes on Science-ing the Humanitie.* http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/ (last accessed on 1 September 2021).
[26] UNESCO Thesaurus. http://vocabularies.unesco.org/browser/thesaurus/en/ (last accessed on 1 September 2021).
[27] Virtual International Authority File. https://viaf.org/ (last accessed on 1 September 2021).

What AI can bring to GLAM? Experience of "Saint George on a Bike" project

Artem Reshetnikov, Barcelona Supercomputing CenterBarcelona, Spain
artem.reshetnikov@bsc.es

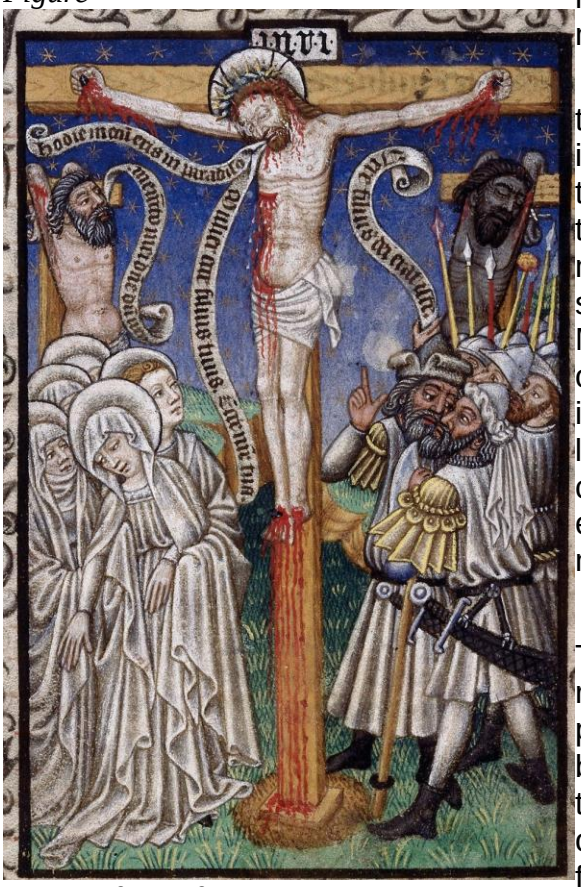
Abstract. *"Saint George on a Bike" project pro- poses several novel approaches to enrichment of meta- data(captions, tags, relationships between objects, iconographic description) for the Cultural Heritage domain, which relies on combining Deep Learning and semantic metadata about paintings. Working with cultural heritage presents challenges not existent for every-day images. Models for objects detection or caption generation are usually trained with datasets that contain correct descriptions of current images or labels for objects, which were generated manually. Apart from this conceptual problem, the paintings are limited in number and represent the same concept in potentially very different styles. Finally, the metadata associated with the images is often poor or inexistent, which makes it hard to properly to generate quality metadata. Our approach can assist in generation of metadata for different tasks. By taking into account an existing metadata of Cultural heritage objects and additional techniques, we can generate tags, relationships between objects or descriptive text which is likely to be directly related to the scene depicted in an image. Index Terms—NLP, Cultural Heritage, Deep Learning, Metadata*

1. Introduction

The application of AI(Artificial Intelligence), and in particular deep learning approaches, to the cultural heritage domain has attracted significant attention in the last time. Most of the existing work focuses on automatic metadata annotation with information such as the author, medium, image classification by style, topic, etc. or the objects that were detected in images from open datasets. However, such types of metadata is not relevant for specific tasks such as generation of descriptions, improving of search engines or improving of communication with users of GLAM sites. Focus of Saint George on a Bike project is on generation metadata, which is related more specifically to cultural heritage domain, which can help to solve these problems. First of all, rich metadata would allow a visitor of a cultural heritage site or the user of a web-page to obtain a detailed description of an artwork and would facilitate a personalized interaction with GLAM institutions. Secondly, different types of metadata could be used to automatically generate explanations in catalogs, fuel search and browse engines, or fill in rich alt-tab descriptions on websites that cater to minorities such as visually impaired citizens. Generating metadata automatically can save a lot of time and labor for manual annotators[1].

The generation of metadata for paintings or images of cultural heritage objects is challenging compared to those corresponding to real world scenes, for several reasons. First, the metadata for paintings often contain irrelevant information beyond the image content such as the life of a historical person, information about the place where the object was found, or the life of the painter. For example, the caption of the artwork in Figure 1

Figure



1: Crucifixion from BL Harley

contains the name of the book where it has been mentioned, the language of the book and the medium of the artwork9.

This information is obviously not relevant for the visual content of the painting. In that con- text, it is challenging to generate good metadata related to the scene. The second challenge is the quality of the data and the data collection process. This makes it difficult to train with a dataset similar in size to datasets containing real life images, such as MS COCO[2]. Lastly, metadata for cultural heritage objects from data providers often contain incomplete sentences or can be in different languages. Data aggregators can't distinguish such cases during data incorporating, as a result, they end up as part of the metadata and affect negatively the quality of the datasets.10

The goal of Saint George on a Bike is to provide rich information about European cultural heritage pictorial artwork. More than one type of output may be generated, which fundamentally depends on the type of input available. The levels of semantic output that we currently contemplate are the following:

Semantic resources in form of tags coming from existing vocabularies

Textual captions

At this point in the project, we have designed and implemented several solutions that can generate textual tags or captions. We have identified the controlled vocabulary from which to choose semantic tags based on the Europeana Entity Collection tags, and we are in the process of:

Refining this vocabulary

9 https://tinyurl.com/ypfbsr66

10 https://tinyurl.com/4rpn6vtf

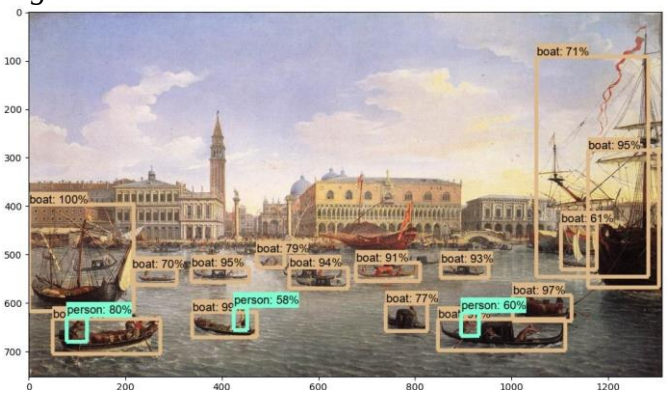
Considering related sources such as DBpedia, Wikidata, or more specific vocabularies used by Europeana providers

In the rest of this document we explain the system and module-level architecture for each of these techniques, as well as their implementation.

2. Object detection

Object detection is a base step for several tasks, including caption generation and search. There are plenty of pretrained models (VGG-16, VGG- 32, ResNet, etc.)[3][4] based on different datasets which can be used in object detection. However, object detection in cultural heritage has its own limitations. These models are usually trained with datasets whose object classes have no symbolic and iconographic dimension. However, when describing paintings, classes cannot be basic and broad-brush. For example, a bishop, Virgin Mary, or Saint George cannot be referred to as just a person when the painting contains object classes that identify them. Even a simple task such as recognizing animals and people can easily convert into a complex task if we'd like to know if the animal is a superbeing (dragon, minotaur, etc.), or what is the occupation of a person. That is why we decided to train our own model using transfer learning, which is able to detect classes with focus on cultural heritage. The detected objects are therefore labeled with our own class names.

Figure



Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. It is a popular approach in deep learning where pretrained models are used as the starting point for computer vision or natural language processing tasks – given the vast compute and time resources required to develop neural network models for these problems and the huge gains it provides when applied to related problems. To improve precision of our object detection model we decided to use transfer learning.

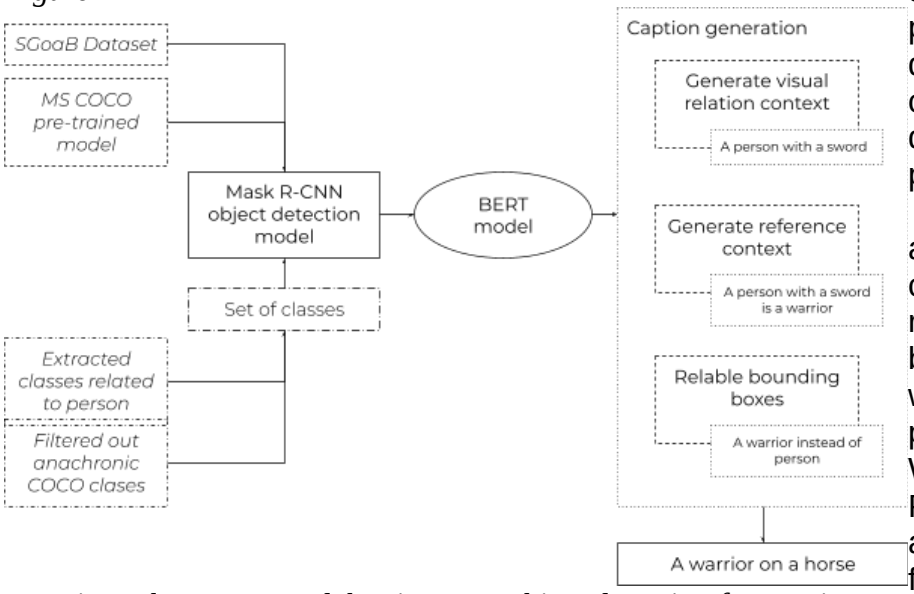
Our implementation uses the Mask-RCNN (Kaiming et al. (2017))[5] model based on the pretrained weights of the MS COCO dataset, as a starting point for the transfer model. The training set consists of more than 13000 manually labeled examples with annotations (source of image, file path, bounding box information, class names) in VOC Pascal XML format (Figure 2). Full list of classes can be find in Appendix. We defined 69 classes based on a careful selection process that first eliminates anachronic classes from the COCO dataset, and then sets to detect the most common objects present in paintings, to further filter this set. A painting class corresponds to a category in a painting collection. The painting collection we have taken as a reference is the Wikimedia Commons collection of paintings labeled by the regular expression "Paintings of. . . "11

11 https://tinyurl.com/26dtr54j

A. Extending the set of classes of interest

The next step extends the set of classes. Wiki- media Commons categories and subcategories are very useful to discern new painting classes when querying about basic classes. For example, if we query about paintings of people we find the subcategory angels_with_humans. In this case, humans is a general reference that covers the basic classes people, men, women and angel is a new painting class because Wikimedia has the category Paintings of people with angels. Starting from the filtered COCO dataset, new classes are added that are related via Wikimedia categories and subcategories. Among the possible classes derived from Wikimedia categories we have chosen a sample with iconographic and symbolic meanings, supernatural and metamorphosed animals (swan in Leda's paintings, cow in the rape of Europa) and devils. Apart from dragons, other fantastic animals are unicorn, centaur, minotaur. We also consider classes that help to identify people that have a social role (occupation) such as bishop, pope, knight or king.

Figure



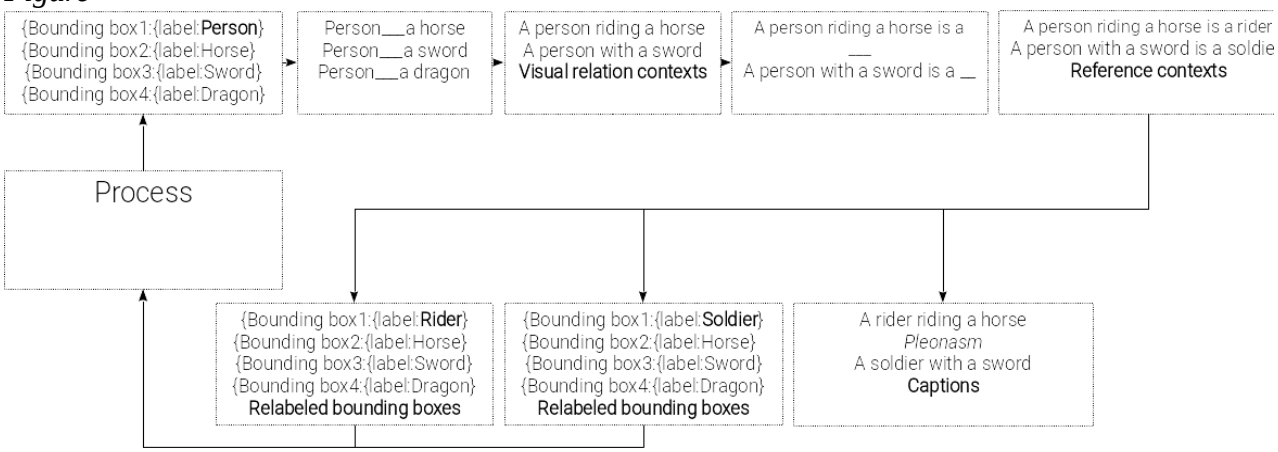
3: Using a language model to improve object detection for caption generation

3. Refining object class detection by using a language model

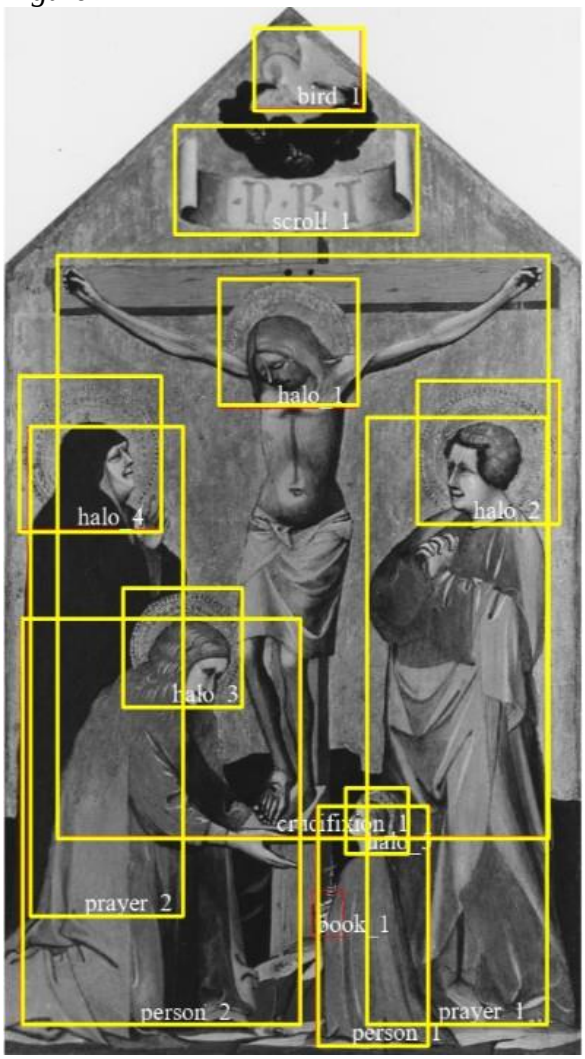
Figure 3 illustrates the caption generation technique that we have implemented, which is based on a language model (using BERT)[6]. The input to this model is the image representation of the painting containing a set of bounding boxes, one for each object class. The output is a set of statements that explain the visual relationships between the classes in the bounding boxes. In these texts, the classes corresponding to the bounding boxes are referred to with more specific denominations according to their visual relationships.

We use the detection network Mask-RCNN to identify bounding boxes in a painting and generate candidate labels for each of them. Section II explains the transfer learning process

Figure



Figure



5: Detection of relationships between pairs

we apply to train the object detection model, starting from weights provided by a MS COCO pretrained model. The Wikimedia Commons catalog covers iconographic classes (e.g. the Annunciation), symbolic classes (e.g. key of heaven for St. Peter), as well as imaginary beings, occupations (e.g. monks, knights), etc.

The goal of caption generation is to refine the references to the main object(s) among all salient objects of a painting. We consider that the main object is the one whose bounding box intersects the largest number of other bounding boxes in a specific object cluster. For each object whose bounding box overlaps with the main object's, the algorithm first generates a set of sentences that describe the possible visual relations between the two. These texts are generated by using a language model to guess missing words that mask the possible relations between the two objects, and they are called visual relation contexts. We then use the language model again to generate the most appropriate completions that specialize the original object class in the visual relation context. (See figure 4)

For instance, a person carrying a cross is Jesus, while a person with a crown becomes a queen or a king. Only those pairs that are predicted with high accuracy by the language model, will generate a visual relation context.

Each of the visual relation contexts that pass this filter are then placed in a reference context to refine the main object and generate the captions. This tool outputs a set of textual captions and can generate basic level classes, higher-level concepts, and named entities.

4. Visual relationships between detected objects in an image

Another approach of detection of visual relationships between objects is based on analysis of bounding boxes positions. Multiple objects can be successfully detected and labeled in an image (e.g. by R-CNN). However, part of the challenge inherent in building systems for automatic image captioning is that learning the visual relationships between detected objects in an image is not trivial. In this section, we describe how a custom implementation of a bounding-box-based (bbx) analysis yields useful visual relationships between objects previously detected by R-CNN technology. We dubbed this Python based implementation "VIS-REL". Our code is applied to imagery representing sacred art produced between the 14th and the 18th centuries (both included).

The context being that of sacred iconography, producing captions to enrich image annotations is a task that broadly corresponds to Panofsky's second level of interpretation of cultural heritage imagery. An example of that would be for the image beholder or the image processing system to rightfully conclude that 13 men having supper with bread and wine (primary level of interpretation) represent the figure of Jesus Christ flanked by his 12 apostles in "The Last Supper" before his crucifixion in Jerusalem (secondary level of interpretation) as described in the New Testament of the Christian Bible. The general idea of the approach is based on detection of relationships between pairs of objects. In order to assess whether any two detected objects, belonging to any two arbitrary object classes (e.g. a person and a horse), are in the same image view-plane, that is to say, at the same field depth in an image, one needs a base-reference of pairwise proportions between objects of every trained class (Figure5).

In practice, VIS_REL computes pairwise- proportions based on common-sense measures and proportions translated as relative surface area proportions between bbxes. Those pairwise proportions between detectable objects are meant to reflect a common-sense representation of realistic pictorial proportions in paintings. Comparing of proportions and some additional measurements allows defining rules which can assume general relationships between pairwise objects:

- Stands
- Holds
- Sits
- On
- Behind
- etc.

5. Challenges

Despite the progress of the project, the technology remains significantly more primitive than human vision and cannot yet satisfactorily ad- dress all challenges of GLAM-institutions. We see a number of long-standing challenges:

- Data collection
  - Some classes are represented only in a few images
  - Style, medium, color may differ significantly between artists

  - Not so many paintings anyway and can't produce them when needed
- Poor metadata
  - Labeled bounding boxes
  - Descriptions of visual content
  - Labeled visual relationships
- Small dataset of paintings by data mining standards requires complementary techniques to
  - Filter out anachronisms
  - Detect imaginary objects or (unusual) actions
- Evaluation
  - Quantifying enrichments quality and usefulness to the user

6. Future work

Future work is structured around several directions:

- Improve the current methods for tag and caption generation
  - Increase training dataset size for object detection to about 15k pictures. This is the set that includes bounding box information and not the image/caption pairs dataset.
  - Use our own trained model (described in Section "Object detection") as an encoder for caption generation using the attention mechanism.
  - Collect training dataset size for caption generation with relevant canonical captions that can be effectively analyzed via Natural Language Processing techniques.
  - Look into the evolving meanings of a word, or homonymic meanings of words, to be able to deal with different meanings over (potentially) distinct time intervals.
  - Test other language models besides standard BERT (e.g. EuBERT). Other approaches to caption classification may be possible, such as fitting a language model over the image/caption dataset.
- Source more data for training and/or evaluation, notably by crowdsourcing.
- Update processes in Section III, so that resulting textual tags are 'uplifted' to semantic tags.
- Build a knowledge graph for the domain of expertise. Complement the work on inferring visual relationships based on BBx' analysis with an approach that could start from knowledge graphs and domain axioms and refine or infer richer object labels and relationship names. These will translate in the generation of semantic graphs for the images. This task involves a thorough evaluation step, the result of which will determine the ability to generate good metadata about basic and higher level actions.
- Extend the scope of the methods to more general topics beyond figurative and mostly iconographic paintings

7. Acknowledgment

This research has been supported by the Saint George on a Bike project 2018-EU-IA-0104, co- financed by the Connecting Europe Facility of the European Union.

8. References

[1] Shurong Sheng, Marie-Francine Moens.2019. ”Generating Captions for Images of Ancient Artworks”. The 27th ACM International Conference

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Doll´ar.2014.”Microsoft COCO: Common Objects in Con- text”. arXiv: 1405.0312.

[3] Karen Simonyan, Andrew Zisserman.2014. ”Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv:1409.1556 .

[4] Karen Simonyan, Andrew Zisserman.2015. ”Very Deep Convolutional Networks for Large-Scale Image Recognition”. arXiv:1409.1556

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. 2015. ”Deep Residual Learning for Image Recognition”. CVPR

[6] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, Ross Gir- shick. 2017. ”Mask R-CNN”. ICCV

[7] YJacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.2019. ”BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding”. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Apendix A

List of classes

Crucifixion, Angel, Person, Crown of thorns, Horse, Dragon, Bird, Dog, Boat, Cat, Book, Sheep, Shepherd, Elephant, Zebra, Crown, Tiara, Camauro, Zucchetto, Mitre, Saturno, Skull, Orange, Apple, Banana, Nude, Monk, Lance, Key Of Heaven, Banner, Chalice, Palm, Sword, Rooster, Knight, Scroll, Lily, Horn, Prayer, Tree, Arrow, Crozier, Deer, Devil, Dove, Eagle, Hands, Head, Lion, Serpent, Stole, Trumpet, Judith, Halo, Helmet, Shield, Jug, Holy Shroud, God The Father, Swan, Butterfly, Bear, Centaur, Pegasus, Donkey, Mouse, Monkey, Cow, Unicorn.

Handwritten Text Recognition for the EDT Project. Part I: Model Training and Automatic Transcription*

Joan Andreu Sánchez and Enrique Vidal

PRHLT, Uinversitat Politècnica de València (UPV) *and* tranSkriptorium AI S.L. (tS)

(jandreu,evidal)@transkriptorium.com

Abstract

Many massive handwritten text document collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried behind thousands of terabytes of high-resolution images. If perfect or sufficiently accurate text image transcripts were available, image textual content could be straightforwardly indexed for plain-text textual access using conventional information retrieval systems. But fully automatic transcription results generally lack the level of accuracy needed for reliable text indexing and search purposes. And manual or even computer-assisted transcription is entirely prohibitive to deal with the massive image collections which are typically considered for indexing. This paper explains how very accurate indexing and search can be directly implemented on the images themselves, without explicitly resorting to image transcripts. Results obtained using the proposed techniques on several relevant historical data sets are presented, which clearly support the high interest of these technologies.

1 Introduction

In recent years, massive quantities of historical handwritten documents are being scanned into digital images which are then made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive text *image* collections are available through Internet. The interest of these efforts not withstanding, unfortunately these document images are largely useless for their primary purpose; namely, exploiting the wealth of information conveyed by the text captured in the document images. Therefore, there is a fast growing interest in automated methods which allow the users to search for the relevant textual information contained in these images which is required for their needs.

In this sense, the project "European Digital Treasures (EDT): Management of centennial archives in the 21st century" [1] aims at bringing joint European heritage, especially its digital versions, major visibility, outreach and use. Three main goals are defined in the EDT project:

- To conceptualize and generate new business models that seek the profitability and economic sustainability of the digitized heritage of archives.
- To foster the development of new audiences especially focused on two groups: the young and the elderly - the latter so-called "golden-agers" or the "silver generation" made up of retirees and citizens aged 60+.
- To promote the transnational mobility of managers, historians, experts, graphic artists, industrial designers and archivists, working on the production of new technological products and interactive exhibitions that give support and visibility to three major European cultural areas.

*The second part of this publication deals with textual information search in untranscribed manuscripts and will appear in (Vidal and Sánchez, 2021).

[1] https://www.digitaltreasures.eu

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image textual content could be straightforwardly indexed for plain-text textual access. However, OCR technology is completely useless for typical handwritten text images; and fully automatic transcription results obtained using state-of-the art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes(Graves et al., 2009; Romero et al., 2012*a*; Vinciarelli et al., 2004).

An alternative to fully automatic processing is to rely on *computer-assisted* transcription. This was successfully explored empirically in (Alabau et al., 2014; Romero et al., 2012*b*; Toselli et al., 2010), following new, powerful concepts of pattern recognition-based human-machine interaction introduced in (Vidal et al., 2007) and (Toselli et al., 2011). In the last eight years, the TRANSCRIPTORIUM and READ projects[2], have further explored the capabilities of these automatic and interactive HTR (IHTR) technologies to speed-up the conversion of raw text images into electronic text.

Working conclusions from all these studies are as follows:

a) To some extent, fully automatic transcripts of text images can be useful for plain-text indexing and search purposes. However, in many historical text image collections of interest, the typical level of transcription accuracy achieved severely hinders the search *recall*; i.e., the system ability to ensure that all or most of the images where a given query text appears can actually be retrieved.

b) Similarly, the fully automatic transcription of most historical text images do not reach the level of accuracy needed for typical scholarly editions of the corresponding image collections.

c) In both cases, the required level of accuracy can obviously be obtained by means of additional user effort. If human work is to be done, rather than just letting the users to edit the noisy automatic transcripts, IHTR can be used to cost-effectively provide the desired transcription accuracy.

e) IHTR can lead to significant gains in human effort with respect to just manually editing the automatic transcripts. But the overall human effort demanded by IHTR is still substantial. Therefore, while IHTR is proving useful to produce scholarly editions of moderately sized historical collections, the required effort to deal with the kind of massive image collections, which are the typical target of indexing and search, is by all means entirely prohibitive.

Given the current situation of the HTR technology that have previously been described, in the last decade the Probabilistic Indexing (PrIx) technique (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Toselli et al., 2019; Vidal et al., 2020) has emerged as a solid technique for making the searching in document images a reality. This technique provides a nice *trade-off* between the *recall* and the *precision* that allows the user to locate most of the relevant information that s/he is looking for in large image collections. The techonolgy is introduced in (Vidal and Sánchez, 2021), mainly focused in the EDT collections.

The approaches proposed here are training-based and therefore need some amount (tens to hundreds) of manually transcribed images to train the required optical and language models. In addition they may benefit from the availability of collection-dependent lexicon and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources will be more than rewarded by the benefits of accurately making the textual contents of these collections available for exploration and retrieval.

The proposed HTR and PrIx approaches have been tested in the past for many historical collections of handwritten text images. Most of the early work was carried out within the TRANSCRIPTORIUM and READ projects mentioned above. The results of these experiments can be seen in a number of recent publications[3]. Here we will present new experiments carried out with manuscript collections researched in the EDT project.

[2] `http://transcriptorium.eu`, `http://read.transkribus.eu`

[3] See: (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Sánchez et al., 2019; Toselli et al., 2019; Vidal et al., 2020)

2 Preparation of a HTR System

Off-line automatic Handwritten Text Recognition (HTR) is a challenging problem that requires a careful combination of several advanced Pattern Recognition techniques, including but not limited to Image Processing, Document Image Analysis, Feature Extraction, Neural Network approaches and Language Modeling.

HTR has progressed enormously in the last two decades due mainly to two reasons: first, the use of holistic training and recognition concepts and techniques which were previously developed in the field of Automatic Speech Recognition (ASR); and second, the existence of an increasing number of publicly available datasets for training and testing the HTR systems.

The need for holistic techniques in HTR has been known for many years given that the processes of handwriting and speech share many similar properties and challenges (Bazzi et al., 1999; Graves et al., 2009; Toselli et al., 2004): i) in both cases the production process is sequential through time; ii) the resulting images or signals are often largely distorted and severely contaminated with different kinds of noise; iii) due to the sequential production process, it is not possible in general to accurately recognize isolated words or characters/phonemes because none of these units can be reliably and consistently segmented or isolated; and iv) handwriting images and speech signals typically exhibit similar forms of lexical and syntactical regularity and ambiguity. Because of these similarities it is not surprising that the same basic Pattern Recognition techniques which had proved successful in ASR also become successful in HTR. To name a few: hidden Markov models (HMM) and recurrent neural networks (RNN) for optical character/phoneme modeling and statistical N-gram models for language modeling. These models are trained both in ASR and HTR with identical machine learning techniques based on the use of annotated data. The availability of sufficiently large amounts of annotated data is currently one of the bottlenecks to move forward in HTR since the annotation is generally performed by human experts and is, therefore, expensive and time-consuming.

The most traditional approaches to HTR are based on N-gram language models (LM) and optical modeling of characters by means of HMMs with Gaussian mixture emission distributions (HMM-GMM) (Marti and Bunke, 2001). In the last decade, notable improvements in HTR accuracy have been achieved by using RNNs for optical modeling. As of now, the state-of-the-art optical modeling HTR technology is based on deeply layered neural network models (Bluche, 2015; Bluche et al., 2015; Graves et al., 2009). The overall architecture is often referred to as *Convolutional-Recurrent Neural Networks* (CRNN) (Shi et al., 2015).

Trained N-grams are represented as a stochastic finite-state transducer. The stochastic transducer, along with the classical Viterbi decoding algorithm (also known as "token- or message-passing"), are used to obtain an optimal transcription hypothesis of the original input line image.

The training process of the optical models is performed both with line images and their corresponding transcripts that need to be in correspondence. Consequently both layout and transcripts have to be prepared to train the HTR system. Layout annotation is related to the mark-up of relevant regions (text regions, marginalia, images, plots, etc.) and the mark-up of (base-)lines located in these regions. In this paper we assume that this process is performed only to get the baselines. Related to the transcript of the lines, they are are referred as "ground-truth" (GT) and they are usally prepared by experts. This GT preparation process is very crucial and relevant decissions have to be made that can affect further steps. Some of the decissions to be made are:

- To use or not to use modernized transcripts.
- To use or not to use all transcripts in capital letters or to use mixed case.
- To expand or not to expand abbreviated words.
- To use or not to use semantic tags.
- To use or not to use diacritics.
- To annotate or not to annotate hyphenated words.
- How to deal with dates and other figures (ages, temperatures, etc).
- To distiguish or not to distiguish between printed and handwritten characters.

The decissions on these points depends on the final goal of the HTR process and can affect the final results. The GT preparation can be more or less expensive depending on the made decissions. For example,

the easier and less time consuming GT production could be to transcribe line images by using modernized transcripts, in capital letters, with all abbreviation expanded, without semantic tags, without diacritics, without annotating hyphenated words, with all dates in the same format. It is important to remark that the more richer transcripts are the more GT data is necessary and, consequently, more expensive.

In the case of the EDT project, some collections were annotated with tags and other not. The purpose of these tags was to make easier to locate words according to their semantics. We describe in more detail the annotation process for the Spain collection although other collections were annotated in a similar way.

The Spain dataset that was used to prepare the HTR system was annotated with several tags. Each transcript was processed as Figure 1 shows. The HTR model is able to learn aligning each character in the image with a character in the transcript. This figure shows that some characters (<print>, <age>) do not have a visual representation, but the optical model is able to capture some contextual information from the surrounding text. The LM also helps in the recognition process of the tags. Note also that the image contains the abbreviated forms “sol” for “soltera” (femenine form for single) and “lab” for “labores” (housekeeper).

De<print> 36<age> años<print> Estado<print> soltera<civilstate> Profesión<print> labores<job>

Figure 1: Example of a text line used for training. Both the line image (top) and the corresponding tagged transcript (bottom) are needed.

This tagging scheme lets the system learn to distinguish identical words that are used with different semantic meaning; for instance, “Juan<surname>”, “Juan<name>”, “Juan<place>”, “Juan<residence>”.

It is important to remark that the transcripts prepared for the GT are both used to train the optical model and the language model. With enough context, both the optical model and the language model are able to take into account the tags and they can be hypothesised in the recognition process.

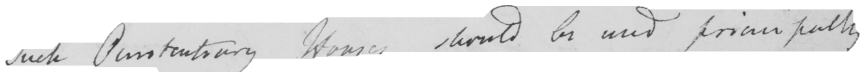
The HTR trained system is used to obtain the most probable transcript by using the Viterbi algorithm. As previously commented, the optical model and the language model are combined in a stochastic finite-state transducer. The best hypothesis uses to lack the level of accuracy needed for reliable text indexing and search purposes. An alternative solution is to obtain the N-best hypothesis transcriptions associated to a line image and to collapse them into a Word Graph (or Character Graph). When N is sufficiently large the Word Graph (WG) associated to a line is able to generalize and to include alternative solutions that were not included in the list on N best hypotheses (Toselli et al., 2016). These WG have recently used to obtain word distributions for each page image rather than just one hypotheses per line image. This idea is explained in (Vidal and Sánchez, 2021).

3 Evaluation Metrics

The most usual evaluation metrics for measuring the performance of an HTR system are the Word Error Rate (WER) and the Character Error Rate (CER). WER is defined as the minimum number of words that need to be substituted, deleted, or inserted to match the recognition output with the corresponding reference ground truth, divided by the total number of words in the reference transcripts. CER is defined in the same way but at character level. See examples in Figure 2.

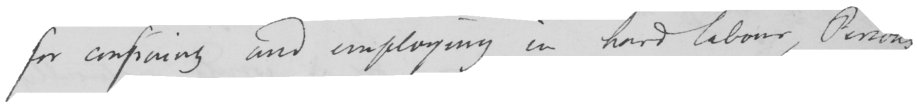
Generally speaking, WER is fairly well correlated with CER, but this correlation is not always strong or systematic. Therefore, *both* measures are important and complementary to assess the quality of an automatic transcript. A low CER but a relatively high WER reveals that the character errors are spread among many words. Conversely, a transcript with the same CER as before but lower WER indicates that errors are concentrated in few words. A good language model typically helps to achieve greater improvements in WER than in CER.

WER tends to be better than CER at indicating how difficult is to understand a transcript by human beings. Similarly, even if CER is low, a high WER may dramatically harm the performance of information extraction or



such Penstentrary Hoases should be anid priapalty

WER = 4/7 = 57%
CER = 8/50 = 16%



for eomfromiy and employing in hard lebour , Persons

WER = 2/9 = 22%
CER = 8/52 = 15%

Figure 2: Two examples of test line images and *automatic* transcripts, along with the corresponding WER and CER. While CER is similar in both transcripts, that with higher WER may be harder to understand. *Reference* transcripts for the top and bottom line images are, respectively: "`such Penitentiary Houses should be and principally`" and "`for confining and employing in hard labour , Persons`".

searching systems which rely on automatic transcripts. Figure 2 illustrates these facts for two samples which exhibit similar CER but different WER.

4 Datasets

Here we will present the datasets compiled in the EDT project for the experiments and results to be presented in Sec. 5. It is important to remark that collections that have been processed in the EDT project are composed by thousands of images but here we focus only in the sets that have been used for preparing the HTR systems.

EDT Hungary. This collection is composed of table images that contain Hungarian proper names in a left-most image region. Figure 3 shows examples of these images. This region is automatically detected by layout analysis techniques and only text lines in this region are detected and extracted. Significant difficulties of this collection include:

- Layout: these tables have a specific but fairly regular layout. They contain both printed and handwritten text, but most cells are typically empty. Only the proper names located in the left-most image region were interesting in the EDT project. Each proper names is preceeded by a handwritten number that had to be avoid in the line detection process to avoid recognition problems.
- Optical modeling: words contain many diacritics and the text has been written by very many hands. This leads to a very high writing style variability.
- Language modeling: the handwritten text are mainly proper names and most of the them are abbreviated in a not consistent way.

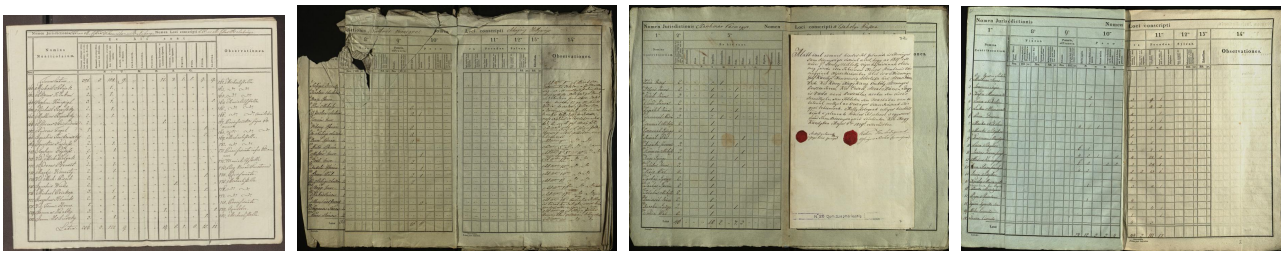


Figure 3: Examples of text images from the EDT Hungary collection.

For the ground-truth preparation, the left columns was annotated with a rectangular region. The main figures of the dataset that was used for preparing the HTR system are reported in Table 1. For training, the lines were

Table 1: Main values of the EDT Hungary dataset.

	Train	Validation	Test	Total GT
Images				410
Lines	7 000	800	672	8 472
Vocabulary	4 168	770	657	4 768
Character set				74
Running words	14 000	1 684	1 403	17 687
Running chars.	119 667	13 804	11 396	144 867

shuffled at collection level and therefore lines from the same pages may be included in all GT partitions. A list of proper names was also provided by the archive that was used to improve the training of the LM.

EDT Norway. This dataset is composed of images of register cards that contain mainly Norwegian proper names. Figure 4 shows examples of these images. Significant difficulties of this collection include:

- Layout: these cards have a fairly regular layout, but it is fairly complex. Both printed an handwritten text is considered for recognition. Each card contains record space for two persons, which appear in separate (left and right) image regions. However, handwritten information can be given for just one person or for both. Most card fields are typically empty and some stamps appear in many cards.
- Optical modeling: some special characters and diacritics are used and the cards have been filled by very many hands. This leads to very high writing style variability, and more so for the size of the initial capitals, which tends to be much larger than the main text body.
- Language modeling: the handwritten text contains many proper names and dates. Many date formats are used, but the archive wished to handle dates in a standard format. Only selected information items of the cards were interesting for the archive.

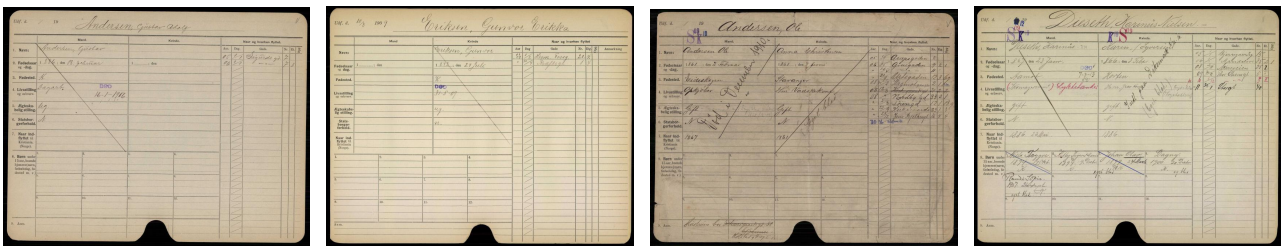


Figure 4: Examples of text images from the EDT Norway collection.

For the ground-truth preparation, the left column of each card was annotated with a rectangular region. Then, the space for two persons were also annotated with rectangular regions along with the proper name in the header. This made easier the training of a specific layout analysis system. This made easier the training of a specific layout analysis system. Lines were detected only inside these regions. The main figures of this dataset are reported in Table 2.

For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions. Printed and handwritten characters were modelled without distinction.

EDT Portugal. This dataset is composed of grayscale images with running text written in Portuguese. Figure 5 shows examples of these images. Significant difficulties of this collection include:

Table 2: Main values of the EDT Norway dataset.

	Train	Validation	Test	Total GT
Images				36
Lines	5 000	243	200	5 443
Vocabulary	1 588	180	131	1 676
Character set				74
Running words	12 069	581	328	13 100
Running chars.	65 405	3 195	2 479	71 079

- Layout: baseline detection is difficult because text lines generally exhibit a great amount of warping, along with very variable slopes and slant. Layout becomes often complex because of plenty marginalia and other more complex layout structures. Many images include parts of adjacent pages.
- Optical modeling: Text include many diacritics and has been written by several hands leading to a significant amount of writing style variability.
- Language modeling: The text contains many proper names and dates and a greet amount of abbreviations and hyphenated words.

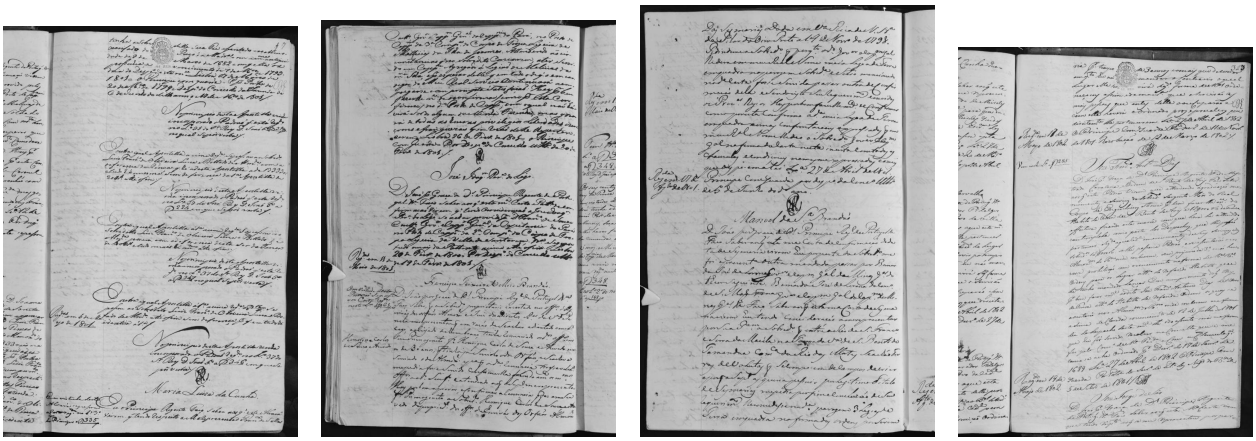


Figure 5: Examples of text images from the EDT Portugal collection.

The main figures of this dataset that are used for preparing the HTR system are reported in Table 3.

For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions. It is worth mentioning that the amount of handwritten text for training is reasonable, as it surpasses the desirable amount of at least 10K words.

Table 3: Main features of the EDT Portugal dataset.

	Train	Validation	Test	Total GT
Images				36
Lines	1 200	122	92	1 414
Vocabulary	2 034	526	424	2 281
Character set				78
Running words	11 338	1 131	816	13 285
Running chars.	62 785	6 359	4 685	73 829

EDT Spain. This dataset is composed of images that contain visa records of Spanish citizens for traveling worldwide. They were issued between years 1936 and 1939 in a Spanish consulate based in Buenos Aires (Argentina). Figure 6 shows examples of these images. Significant difficulties of this collection include:

- Layout: Each image contains four visa forms and each form includes one (or several) picture(s) associated to each visa. Some dates are stamped and others are handwritten. Text lines often exhibit extreme slope.
- Optical modeling: Forms include printed text and are filled with handwritten text written by many hands, leading to a very high writing style variability.
- Language modeling: the handwritten text contains many proper names (given names, surnames, town and state names, countries, etc.) and dates. A large amount of words are heavily abbreviated. All the textual information in each visa was relevant for the archive.

Figure 6: Examples of text images from the EDT Spain collection.

A small set of 99 images of the whole collection were selected for ground-truth annotation. For layout analysis, each image was annotated with four rectangular regions to isolate each visa, and then the geometric data of each photograph region and all the baselines were annotated. Then each line was manually transcribed and annotated word by word with “semantic” tags.

The tags that have been used in this dataset and their meanings are:

- `<print>`: printed word of the form
- `<date>`: date, both stamped or handwritten
- `<gname>`: given name
- `<surname>`: surname (two surnames are used in Spanish)
- `<state>`: province
- `<country>`: country
- `<civilstate>`: civil state (single, married, etc.)
- `<residence>`: place of residence
- `<place>`: location (city, village, etc.)
- `<job>`: occupation
- `<age>`: years old

The main values of this dataset are reported in Table 4. Since the amount of handwritten text and printed text can significantly affect the recognition results, this table provides the amount of both types of text. Experiments will be reported without taking into account this difference. Since the tags are considered special characters, the same word with two different tags was considered two different words. This fact explains the large vocabulary that can be observed in the “Hand” columns.

For the training, the lines were shuffled at collection level and therefore lines from the same pages may be included in the training and in the test partitions. It is worth mentioning that the amount of handwritten text for training is reasonable, as it surpasses the desirable amount of at least 10K words.

Table 4: Main features of the EDT Spanish dataset. The vocabulary is shown both with and without numbers

	Train		Validation		Test		Total GT	
	Print	Hand.	Print	Hand.	Print	Hand.	Print	Hand.
Images							99	
Lines	6 500		250		261		7 011	
Vocabulary w numbers	317	2 900	48	254	49	276	333	3 057
Vocabulary w/o numbers	53	2 040	36	203	37	202	53	2 149
Character set							85	
Running words	14 662	10 530	575	390	602	455	15 839	11 375
Running chars.	108 226	72 019	4 329	2 585	4 503	3 046	117 058	77 650

EDT Malta. This collection is composed of grayscale images with lists of proper names and the name of the flight or the ship in which the people arrived to Malta Figure 7 shows examples of these images. Significant difficulties of this collection include:

- Layout: the lines have quite slope, slant, skew and warping. Many images are physically degraded with many holes because of woodwarm and other insects. Many images include parts of adjacent pages.
- Optical modelling: there are fainted text and several hands, and many flights and ship names are replaced with quotes (“).
- Language modeling: the text contains mainly proper names and dates.

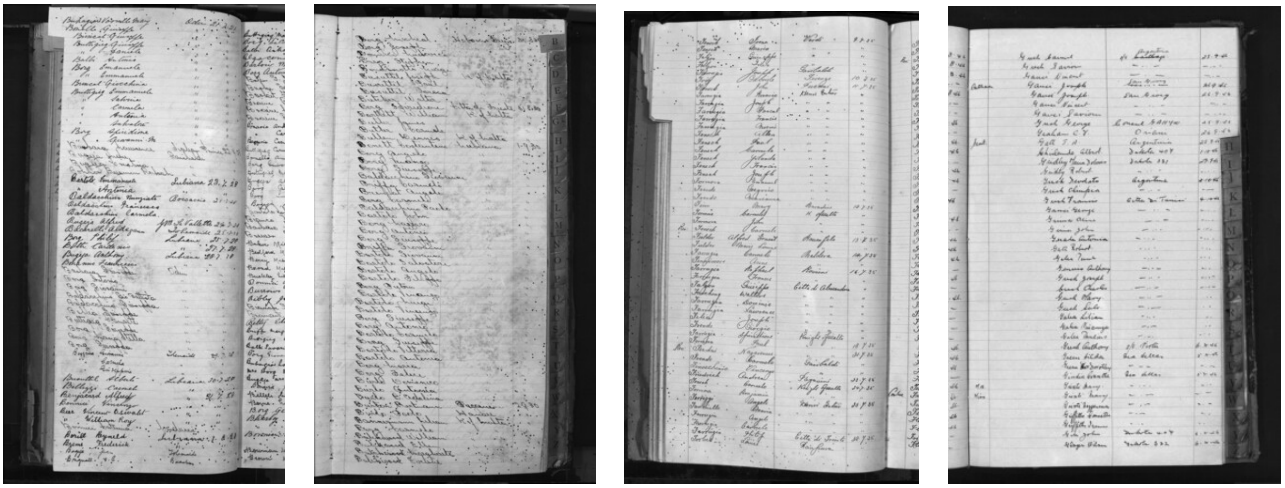


Figure 7: Examples of text images from the EDT Malta collection.

The main figures of this dataset used for preparing the HTR system are reported in Table 5. For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions.

Table 5: Main features of the EDT Malta dataset.

	Train	Validation	Test	Total GT
Images				49
Lines	2 200	230	101	2 531
Vocabulary	3 249	551	274	3 614
Character set				77
Running words	9 996	1 037	448	11 481
Running chars.	65 376	6 829	2 971	75 176

5 Experiments and Results

The datasets introduced in Section 4 were used to prepare an HTR system for each collection. Theses HTR systems were evaluated using the partitions described for each collection. CER and WER were computed and the results are shown in Table 6. Several values of N were used for the N-gram LM but only the results obtained with $N = 5$ are shown since this was the best value or other values of $N > 5$ did not get a significant improvements.

Table 6: CER and WER obtained for each collections.

Dataset	CER	WER
Hungary	8.8	25.7
Norway	5.6	13.4
Portugal	14.5	35.4
Spain	9.4	20.8
Malta	12.8	36.6

We observe that both CER and WER have large difference among collections. The worst result are obtained for the Portugal and Malta collections. In the case of the Portugal collection, the bad WER results are due mainly to the large amount of abbreviated forms. This WER could be decreased by using additional GT. In the case of the Malta collection, the problem is again with the amount of training data, data shoul clearly increased.

6 Conclusion and outlook

We have introduced experimental results for the EDT collections. The obtained results reveal that additional GT data should be prepared for improving the training of the models. For future work, we expect to add additional training material produced in the EDT project with a crowdsourcing initiative.

7 Acknowledgments

References

Alabau, V., Martínez-Hinarejos, C., Romero, V. and Lagarda, A. (2014), 'An iterative multimodal framework for the transcription of handwritten historical documents', *Pattern Recognition Letters* **35**, 195–203. Frontiers in Handwriting Processing.

Bazzi, I., Schwartz, R. and Makhoul, J. (1999), 'An omnifont open-vocabulary OCR system for English and Arabic', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(6), 495–504.

Bluche, T. (2015), Deep Neural Networks for Large Vocabulary Handwritten Text Recognition, PhD thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur. Discipline : Informatique.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H. and Vidal, E. (2017), Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project, *in* 'Int. Conf. on Document Analysis and Recognition (ICDAR)', Vol. 01, pp. 311–316.

Bluche, T., Ney, H. and Kermorvant, C. (2015), The LIMSI/A2iA Handwriting Recognition Systems for the HTRtS Contest, *in* 'International Conference on Document Analysis and Recognition', pp. 448–452.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009), 'A Novel Connectionist System for Unconstrained Handwriting Recognition', *IEEE Transaction on Pattern Analysis and Machine Intelligence* **31**(5), 855–868.

Lang, E., Puigcerver, J., Toselli, A. H. and Vidal, E. (2018), Probabilistic indexing and search for information extraction on handwritten german parish records, *in* '2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)', pp. 44–49.

Marti, U.-V. and Bunke, H. (2001), 'Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system', *International Journal of Pattern Recognition and Artificial Intelligence* **15**(1), 65–90.

Puigcerver, J. (2018), A Probabilistic Formulation of Keyword Spotting, PhD thesis, Univ. Politècnica de València.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2020), Advances in handwritten keyword indexing and search technologies, *in* A. Fischer, M. Liwicki and R. Ingold, eds, 'Handwritten Historical Document Analysis, Recognition, And Retrieval-State Of The Art And Future Trends', Vol. 89, World Scientific, pp. 175–193.

Romero, V., Toselli, A. H. and Vidal, E. (2012*a*), *Multimodal Interactive Handwritten Text Transcription*, Series in Machine Perception and Artificial Intelligence (MPAI), World Scientific Publishing.

Romero, V., Toselli, A. and Vidal, E. (2012*b*), *Multimodal Interactive Handwritten Text Recognition*, Vol. 80 of *Machine Perception and Artificial Intelligence*, World Scientific.

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M. and Vidal, E. (2019), 'A set of benchmarks for handwritten text recognition on historical documents', *Pattern Recognition* **94**, 122–134.

Shi, B., Bai, X. and Yao, C. (2015), 'An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition', *CoRR* **abs/1507.05717**.

Toselli, A. H., Juan, A., Keysers, D., González, J., Salvador, I., Ney, H., Vidal, E. and Casacuberta, F. (2004), 'Integrated handwriting recognition and interpretation using finite-state models', *International Journal of Pattern Recognition and Artificial Intelligence* **18**(4), 519–539.

Toselli, A. H., Romero, V., Vidal, E. and Sánchez, J. A. (2019), Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, *in* '15th Int. Conf. on Document Analysis and Recognition (ICDAR)'.

Toselli, A. H., Vidal, E., Romero, V. and Frinken, V. (2016), 'HMM word graph based keyword spotting in handwritten document images', *Information Sciences* **370-371**, 497–518. Information Sciences 370-371 (2016) 497-518.

Toselli, A., Romero, V., i Gadea, M. P. and Vidal, E. (2010), 'Multimodal interactive transcription of text images', *Pattern Recognition* **43**(5), 1814–1825.

Toselli, A., Vidal, E. and Casacuberta, F. (2011), *Multimodal Interactive Pattern Recognition and Applications*, 1st edition edn, Springer.

Vidal, E., Rodríguez, L., Casacuberta, F. and García-Varea, I. (2007), Interactive pattern recognition, *in* 'International Workshop on Machine Learning for Multimodal Interaction', Springer, pp. 60–71.

Vidal, E., Romero, V., Toselli, A. H., Sánchez, J. A., Bosch, V., Quirós, L., Benedí, J. M., Prieto, J. R., Pastor, M., Casacuberta, F., Alonso, C., García, C., Márquez, L. and Orcero, C. (2020), The carabela project and manuscript collection: Large-scale probabilistic indexing and content-based classification, *in* '17th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)', pp. 85–90.

Vidal, E. and Sánchez, J. A. (2021), Handwritten text recognition for the EDT project. Part II: Textual information search in untranscribed manuscripts, *in* M. A. Bermejo et al., ed., 'Proc. of the EDT Alicante workshop', To appear.

Vinciarelli, A., Bengio, S. and Bunke, H. (2004), 'Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 709–720.

Handwritten Text Recognition for the EDT Project. Part II: Textual Information Search in Untranscribed Manuscripts*

Enrique Vidal and Joan Andreu Sánchez

PRHLT, Uinversitat Politècnica de València (UPV) *and* tranSkriptorium AI S.L. (tS)

(evidal,jandreu)@transkriptorium.com

Abstract

Many massive handwritten text document collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried behind thousands of terabytes of high-resolution images. If perfect or sufficiently accurate text image transcripts were available, image textual content could be strightforwardly indexed for plaintext textual access using conventional information retrieval systems. But fully automatic transcription results generally lack the level of accuracy needed for reliable text indexing and search purposes. And manual or even computer-assited transcription is entierely prohibitive to deal with the massive image collections which are typically considered for indexing. This paper explains the Probabilistic Indexing technology, which allows very accurate indexing and search to be directly implemented on the images themselves, without explicitly resorting to image transcripts. Results obtained using the proposed tecniques on several relevant historical data sets of the EDT project are presented, which clearly suppport the high interest of these technologies.

1 Introduction

In recent years, massive quantities of historical handwritten documents are being scanned into digital images which are then made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive text *image* collections are available through Internet. The interest of these efforts not withstanding, unfortunately these document images are largely useless for their primary purpose; namely, exploiting the wealth of information conveyed by the text captured in the document images. Therefore, there is a fast growing interest in automated methods which allow the users to search for the relevant textual information contained in these images which is required for their needs.

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image textual content could be strightforwardly indexed for plaintext textual acces. However, as discussed in the first part of this publication (Sánchez and Vidal, 2021), OCR technology is completely useless for typical handwritten text images; and fully automatic, or even computer assited transcription results obtained using state-of-the art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes(Graves et al., 2009; Romero et al., 2012; Vinciarelli et al., 2004).

This situation raises the need of searching approaches specifically designed for large text *image* collections. In these approaches, indexing and search must be directly implemented on the images themselves, without explicitly resorting to image transcripts. On the other hand, rather than "exact" searching (as in plaintext), search has to be performed with a *confidence threshold*, somehow specified by the user as part of the query in order to meet the *precision-recall trade-off* which is considered most adequate in each query.[1]

*The first part of this publication deals with model training and automatic transcription and will appear in (Sánchez and Vidal, 2021).

[1]Depending on the application, confidence thresholds can be specified more or less explicitly. For instance, in cases where the spotting results are provided in the form of ranked lists, the threshold is indirectly defined by the size of the list.

Clearly, such a confidence-based query model can not be properly implemented by just using conventional textual information retrieval methods on the noisy output of an automatic HTR system. Therefore, recognition techniques are needed which attach confidence scores to alternative word recognition hypotheses. Keyword spotting (KWS)[2] is a traditional way to address search problems within this framework. More precisely, KWS aims at determining locations on a text image or image collection which are likely to contain an instance of a queried word, without explicitly transcribing the image(s).

Depending on whether the query is specified by means of an example-image or as just a character string, KWS is respectively qualified as Query-by-Example (QbE) or Query-by-String (QbS). Only the QbS paradigm is conisdered adequate for textual information search in large manuscript collections.

Traditional work on handwritten KWS assumed previous segmentation of the text images into words. However, word pre-segmentation is plainly impossible for millions of historical handwritten images of interest and, even in favorable cases, it is quite prone to errors (Manmatha and Rothfeder, 2005; Papavassiliou et al., 2010), which tend to hinder overall KWS performance significantly (Ball et al., 2006). To overcome this important drawback, recent works[3] assume the (word-unsegmented) *line image* as the lowest search level. This is a convenient setting because, in most cases of interest, text images can be fully automatically segmented into lines with fair accuracy (Bosch et al., 2012; Papavassiliou et al., 2010) and lines are sufficiently precise target image positions for most practical document image search and retrieval applications. Nevertheless, robust line segmentation can also be problematic in many cases and nowadays is considered perhaps the most severe bottleneck to achieve fully automatic processing of handwritten images for KWS and HTR alike. For this reason, our current work aims at indexing full pages, in an attempt to circumvent the need for any kind of image segmentation alltogether.

On the other hand, most of the techniques which have been proposed for KWS can be considered to belong to one of these two broad classes: *training-based* and *training-free*. Training-based KWS methods are generally based on statistical optical (and language) models and typically adopt the QbS paradigm. Conversely, most training-free techniques are based on direct (image) template matching and assume the QbE framework.

The approaches we follow are training-based and therefore need some amount (tens to hundreds) of manually transcribed images to train the required optical and language models. In addition they may benefit from the availability of collection-dependent lexica and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources will be more than rewarded by the benefits of accurately making the textual contents of these collections available for exploration and retrieval.

Traditional KWS technology has aimed at searching for a few (tens, hundreds, or maybe thousands of) *"key words"*, which the users should provide as those that are most interesting to search for information in the considered collection. In many cases, these keywords are assumed to be known beforehand.

Clearly, these assumptions go astray when very large collections of manuscripts are considered. In these cases, users can by no means pre-compile any reasonable list of "keywords", and the only adequate approach is to let the system itself "discover" the words which are likely to appear in the text images.

On the other hand, for very large image collections it becomes computationally unfeasible to build a system that not only process the images and discover likly writen words, but also searches for the arbritary words the users happen to include in their queries. Therefore, the system's work has to be divided in two parts: First, in an *off-line, preprocessing* phase, all likly words are hypothesized and adequately used to index the text images. Then, in an *on-line, search* phase, user's queries are analyzed and the indexed images are searched for the words included in the queries.

We call this approach and the corresponding technologies *Probabilisti Indexing* (PrIx). Results of this apporach for many interesting historic handwritten documents have been published in our recent publications.[4] Here, we will report new results on additional historical collections considered in the EDT project.

[2] See (Cao et al., 2009; Fischer et al., 2012; Frinken et al., 2012; Kamel, 2010; Manmatha et al., 1996; Puigcerver et al., 2016; Rath and Manmatha, 2007; Rodríguez-Serrano and Perronnin, 2009; Toselli and Vidal, 2013*a*; Toselli et al., 2016; Wshah et al., 2012).

[3] See (Fischer et al., 2012; Frinken et al., 2012; Kolcz et al., 2000; Terasawa and Tanaka, 2009; Toselli and Vidal, 2013*a*; Toselli et al., 2016; Wshah et al., 2012).

[4] See (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Toselli, Romero, Vidal and Sánchez, 2019; Vidal et al., 2020).

2 Probabilistic Indexing and Search

An overview of the ideas behind the indexing and search technology we are developping is presented in this section. As previously commented, this technology assumes the *precission-recall trade-off search model* which requires *word confidence scores* computed for adequate regions of the text images of interest.

Pixel-level word confidence scores: the "posteriorgram". A basic concept on which the proposed approach relies is the so called *pixel-level "posteriorgram"*. In a nutshell, it is a probability map computed for a given image X and a possible query word v. At each position (i, j) of X, the posteriorgram provides the posterior probability that the word v is written in some subimage of X which includes the pixel (i, j). Fig. 1 illustrates this concept.

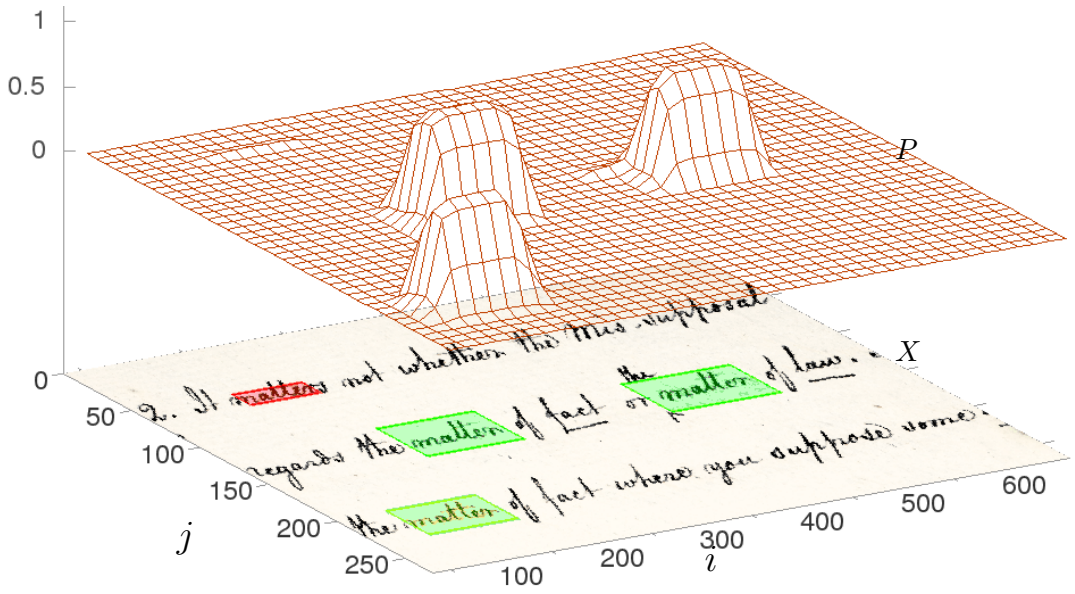


Figure 1: Pixel-level posteriorgram, P, for a text image X and word $v =$"`matter`". The most probable regions of X where v may appear according to P are marked in color boxes (red: low, green: high).

The value of P at each image position (i, j) can be easyly obtained by statistical *marginalization*. In simple words, the idea is to consider that v may have been written in any possible bounding box of the image X which includes the pixel (i, j). The marginalization process simply adds the word recognition probabilities for all these bounding boxes. This means that a posteriorgram could be simply obtained by repeated application of any word classification system capable of recognizing isolated (pre-segmented) words. Obviously, the better the classifier, the better the corresponding posteriorgram estimates.

Directly obtaining a full pixel-level posteriorgram in this way entails a formidable amount of computation. However, as it will be discussed later, it can be very efficiently computed by clever combinations of subsampling of the image positions (i, j) and adequate choices of the marginalization bounding boxes.

In our approaches we use full-fledged holistic HTR systems to compute the required isloated word probabilities. This allows us to take advantage of linguistic context to obtain very accurate word classification probabilities. In Fig. 1, a contextual word classifier based on a n-gram language model was used to compute P for the word "`matter`". This helped to achieve very low probabilities in a region of X around $(i = 100, j = 200)$, where a very similar (but different) word, "`matters`", is written. Clearly, according to the language model, the 2-grams "`the matter`" and "`matter of`" are very likely, thereby boosting the probability that the word "`matter`" is written in the correct image regions. Conversely, the 2-grams "`It matter`" and "`matter not`" are very unlikly, resulting in very low pixel probabilities in the image region where the different word "`matters`" is written (things would roughly be the other way around should the query word be "`matters`" instead).

Image region word confidence scores. Posteriorgrams could be directly used for KWS: Given a confidence threshold τ, a word v is just spotted in all image positions (i, j) where P is greater than τ. Varying the threshold, adequate *precision–recall* tradeoffs can be achieved. However, this naive idea is not feassible for large image collections, simply because indexing word confidences for every image pixel is obviously impossible. For PrIx, what we really need is the confidence that a word v is written within a pre-specified image region, such as a line, a column, or a full page, without explicitly taking into account exactly where the word is written in the region or how many instances of the word may appear in this region. In information retrieval terms, this is called *"relevance"*. That is, for each image region to be indexed, we need to obtain the probability that this region is *relevant* for the given query word.

Exactly computing relevance probabilities can become complex. Nevertheless, a very simple and intuitivelly apealing approach is to obtain the region relevance probability for a word v just as the maximum pixel-level probability for v over all the pixels of the region. For instance, if X in Fig. 1 is considered a region to be indexed, the probability that X is relevant for the the query "`matter`" is adequately approximated just the maximum of the four picks of the posteriorgram shown in this figure.

Choosing adequate minimal searchable image regions: line-level PrIx. In our work so far, line-shaped regions have been adopted as the lowest image element to be indexed. From the user point of view, lines are are sufficiently precise target image positions for most practical document image search and retrieval applications. On the technical ground, on the other hand, line-shaped image regions are particularly useful because they allow for efficient computation of posteriorgrams by adequately choosing the bounding boxes needed for the underlying marginalization process and by clever vertical subsampling of image positions.

Regarding the choice of *marginalization bounding* boxes needed to compute the posteriorgram, for a line-shaped image region, these boxes can be simply defined just by horizontal segmentation.

On the other hand, with line-shaped image regions *vertical subsampling*, in general, amounts to just guessing a proper line height and then runing a vertical-sliding window of this height with some overlap. Moreover, in many cases of interest, text lines are fairly regular and standard line segmentation techniques can yield accurate results. This allows to save computation cost and tends to increase accuracy.

Finally, and most improtantly, line-shaped text image regions typically contain most[5] of the relevant lingusitic context needed for precise computation of word classification probabilities using a language-model based recognizer, as discussed elsewhere.

Efficient computation of posteriorgrams and relevance probabilities. In our approaches, line-level posteriorgrams are very efficiently computed using *Word Graphs*, obtained as a byproduct of recognizing full line-region images with a full-fledged holistic HTR system based on *optical character models* and (N-gram) *Language Models*, as discussed in Part I of this publication (Sánchez and Vidal, 2021). When applied to a line-shaped image region, these systems can take full advantage of the lingusitic context which is present in the image to provide very accurate, word classification probabilities. On the other hand, a WG obtained in this way provides lots of alternative horizontal word-level segmentations. These segments directly define very adequate sets of bounding boxes, exactly as required by the marginalization process used to compute the posteriorgrams.

Line-region relevance probabilities are directly computed from the corresponding posteriorgrams, as explained above. Then, they can in turn be easily and consistently combined to obtain *page-level* relevance probabilities (. . . and so on for higher level indexing of *chapters*, *books*, etc.

Searching for words in probabilistically indexed images. Once the PrIx's of an image collection are available, textual search can be carried out using the pseudo-words and the corresponding geometric information spots contained om the PrIx spots. To this end, classical plaintext search techniques can in principle be applied. It is worth pointing out, however, that since PrIx's are generally very large (as compared with plain text), computing efficiency becomes a major concern.

[5]Most, but not all: linguistic context is obviously lost and the line boundaries. This problem is being considered towards upcoming developments of handwritten search and retrieval technologies.

On the other hand, PrIx easily allows complex queries that go far beyond the typical single keyword queries of traditional KWS. In particular, full support for standard multi-word boolean and word-sequence queries have been developed in (Toselli, Vidal, Puigcerver and Noya-García, 2019) and used in many PrIx search applications for large and huge collections of handwritten text documents.[6]

Most of these applications also provide another classical set of handy free-text search tools; namely, *wild-card* and *approximate* (also called *"fuzzy"* or *"elastic"*) spelling. These tools are generally considered remarkably useful search assets in practice.

Probabilistic Indexing and Search Systems. Fig.2 show a typical textual information retrieval system based on PrIx. Its major components are:

- *"Probabilistic Indexing (PrIx)"*: Off-line pre-computation of PrIx's
- *"Ingest"*: Off-line creation of the actual database. Typically a simple and computationally cheap process
- *"Search engine and GUI"*: On-line user query analisys, find the requested information and present the retrived images.

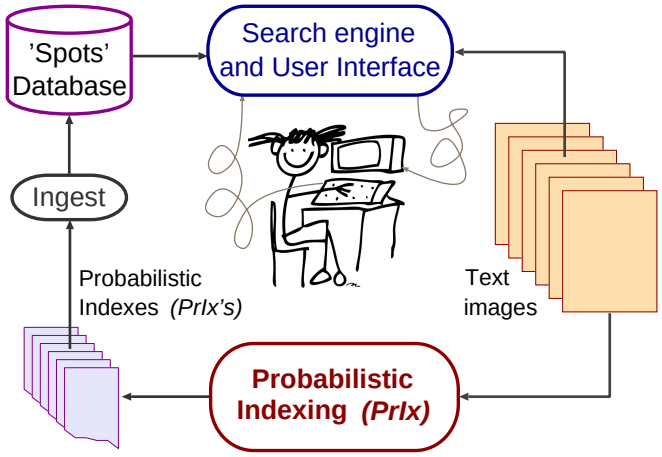


Figure 2: Probabilistic Text Image Indexing and Search System Diagram

PrIx is the most important component. As discussed above, it is based on *contextual word (or char string) recognition*, which requires models *trained* from transcribed images (as in HTR). All this requires heavy (*off-line*) computing – but, as a result, it allows *extremely fast on-line query responses*, even for huge manuscript collections.

Another important component is the search engine and user interface, which should provide:

- *GUI*: Graphical / textual specification of queries and desired precision-recall tradeoff settings;
- *Query analysis*, which is trivial for single words, but becomes complex for multi-word queries, approximate spelling, etc.;
- *Search engine* to Access the database. Specialized software typically needed for probabilistically consistent support of multi-word queries, hierarchical search, and geometry-aware search;
- *Display retrieved images*: Prepare the images to be presented to the users as a result of their queries. The way they are presented is highly application dependent;
- Short response times are needed.

[6] See `http://prhlt-carabela.prhlt.upv.es/PrIxDemos` for a list of PrIx live demonstrators.

3 PrIx Search Evaluation Metrics

The standard *recall* and *interpolated precision* measures (Manning et al., 2008) are used to assess the effectiveness in all the search experiments.

For a given query and confidence threshold, *recall* is the ratio of relevant image regions (lines) correctly retrived by the system (often called "hits"), with respect to the total number of relevant regions existing in the image test set. *Precision*, on the other hand, is the ratio of hits with respect to number of (correctly or incorrectly) retrived regions.

By variing the confidence threshold, different related values of recall and precision can be obtained. These values can be ploted into the so-called *Recall-Precision* curve. Clearly, for a perfect system this curve would go stright from the point $(1, 0)$ vertically up to $(1, 1)$ and then horizontally left to $(0, 1)$. That is, such a system should exhibit a full precison (1) independently of the confidence threshold. This would in fact be the behaviour of a conventional plaintext retrieval system tested on a the perfect transcripts of the test set images. A reassonable search system should provide curves that go above the diagonal of the graph. the closer to the upper right corner (point $(1, 1)$), the better.

Results are also reported in terms of overall *average precision* (AP), which are obtained by computing the area under Recall-Precision curves and is a popular scalar assessment measure in Information Retrieval and KWS alike. Please refer to (Toselli et al., 2016) for details on these assesment measures.

4 Datasets

Many historical collections of handwritten text images have been considered in the past for testing the proposed indexing and search technologies. Most of the early work was carried out within the TRANSCRIPTORIM and READ projects mentioned in Sec. 1 of Part I of this paper (Sánchez and Vidal, 2021). Comprehensive accounts of these experiments have recently been reported in several publications (see footnote 4).

Here we will present new experiments carried out through our collaboration with the EDT project. The features of these datasets are described in Part I of this paper and summarized in Table 1, below. In each of these datasets, the set of query words used in the PrIx search experiments is the full set of words (vocabulary) of the test partition of each dataset.

5 PrIx Search Results

In all the experiments the Optical and Language models used are trained from the training and validation partition of each dataset, as described in Sec. 2 and 6 of Part I of this paper (Sánchez and Vidal, 2021). PrIx search results were assessed using the recall-precision metrics outlined in Sec.3. The average precision for all the datasets is reported in Table 1 and the corresponding R-P curves are shown in Fig. 3.

Table 1: Main EDT dataset features and Average Precision results.

EDT dataset	Hungary	Norway	Portugal	Spain	Malta
Training running words	14 000	12 069	11 338	25 192	9 996
Test running words	1 403	328	817	1 053	448
Training vocabulary	4 168	1 588	2 034	3 217	3 249
Query set size (words)	657	123	417	322	274
Average Precision (AP)	0.76	0.88	0.60	0.77	0.61

Results are reasonably good for all the datasets. Not surprisingly, he amount of training words per vocabulary word (ratio between training running words and training vocabulary size) apears to have a positive impact on the results – clearly, more training data are needed if more different words are expected. Also, performance tends to be inferior for the more difficult collections. But, as discussed below, even at the lower levels of performance the systems can be advantageously used in practice to reliably find relevant information.

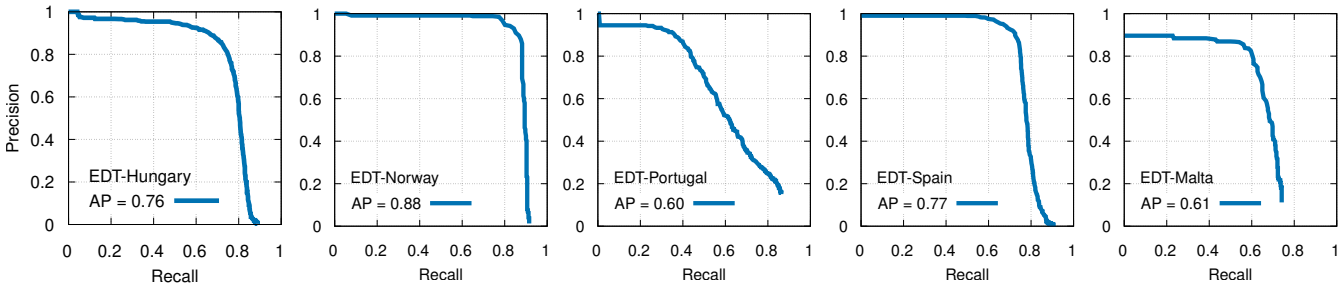


Figure 3: R-P curves for the EDT datasets.

Overall these results are comparable to our previous results using the PrIx technology in other manuscript collections (see footnote 4) and very competitive as compared with results reported in the literature for conventional KWS systems.[7]

However, one may argue that these great laboratory results may not translate into a similarly good practical search experience. Consider for instance searching for information in the EDT–Spain collection. Typically, a user will try to find names of persons or places, or maybe person ages or ocupations. In this scenario, for an R-P operational point such as RECALL $= 0.7$ (and corresponding PRECISION $= 0.93$ – see Fig. 3), about 7% of the retrived spots are expected to be false alarms, which is great – but a 30% of the spots where the query word is actually written could be missed. The user might then try to improve the recall by lowering the confidence threshold and thereby setting another R-P operational point with a lower precision. However, even at a PRECISION as low as 0.1 (90% expected false alarms), the corresponnding RECALL is still 0.84, with an expected number of misses of 16%. Seveal factors, however, tend to make the search experience much better than would be expected according to these numbers.

First, searching for information in handwritten text images can by no means be compared with conventional information retrieval, where no uncertainty exist about the words contained in the (electronic) documents searched for. In handwriten text images, the primary search baseline is just manual search; that is, visually scan each of the (maybe thousands or millions) page images, trying not to miss image regions where the query word may appear. Clearly, even an AP as low as 0.5 or even lower, may prove extreemly useful as compared with the manual baseline.

Second, consider for instance the results of EDT–Portugal dataset, with AP $= 0.60$. These results are averaged over a query set of 452 diferent words, which encompasses all the test–set words, including frequent function words and many other (short, more difficult to spot) words that would probably never be used as part of a query. For typical query words, results are generally better (even thouh specific experiments are needed to validate this assertion).

Finally, it worth to remind that, under the precision-recall tradeof search model, users are not expected to be content with a fixed operational R-P point. Depending on the interest in finding only some, or most of the occurrences of a given query word, users will try increasing or decreasing the confidence threshold until they become satisfied with the results and/or understand they have met the limitations of the system.

The real systems referred to in the next section can be used to get first-hand experience of the capabilities and limitations of these systems and the significance of the results presented in this section.

6 Real PrIx and Search Systems for EDT Collections

Indexing and search engines similar to those used to obtain the results presented in Sec. 5, have also been used to support real search systems implemented according to the scheme of Fig. 2. These systems can be publicly accessed through Internet:

[7] See (Fischer et al., 2012; Frinken et al., 2012; Rath and Manmatha, 2007; Rodríguez-Serrano and Perronnin, 2009; Toselli and Vidal, 2013*b*; Toselli et al., 2016; Wshah et al., 2012)

- EDT-Hungary: `http://edt.transkriptorium.com/hungary-search`
- EDT-Norway: `http://edt.transkriptorium.com/norway`
- EDT-Portugal: `http://edt.transkriptorium.com/por-tr`
- EDT-Spain: `http://edt.transkriptorium.com/esp`
- EDT-Malta: `http://edt.transkriptorium.com/malta`

Note, however, that these system are not exactly the same as those used in the experiments. The most important difference is that in these on-line systems, all the non-essential diacritics and special characters have been ignored and print/handwritten and other "semantic" tags have been removed. In general, this can greatly sinplify the query experience and make it more effective. Nevertheless, for collections such as EDT–Sapain, this actually hinders the system hability to honor complex, semantic-oriented queries and other advanced search capabilites, for which other systems have been set up, [8] though they are not publicly available for the time being.

7 Conclusion and outlook

A formal probabilisitic framework has been introduced for indexing and searching large collections of handwritten documents. Empirical results with a variety of historic collections exhibiting differnet challenges and levels of complexity assess the usefulness of these methods in practice. Several demonstrators have been implemented and made publicly available through the Internet for first-hand experience in real use.

In the coming future, work is planed to adress the folloing issues:

- So far line-regions are considered the most elementary elements to be indexed. This entails a requirement for automatic line detection and extraction. While fairly accurate automatic text line detection techniques exists, results lack robustness; that is, these techniques are not robust enough to reliably deal with the large variability in image quality and layout usually exhibited by historic handwritten documents. So, from time to time, a batch of page images appears in which line detection may fail dramatically. And, as a result, these pages become unindexed. Our current work aims at considering full page images as the lowest indexing level, in an attempt to completely circiunvent the line detection bottleneck (Barrere et al., 2019).
- All the techniques and experiments described in this paper assume that a user query is just a single word. Multiple word combined queries, and more specifically boolean and word sequence combinations (Toselli, Vidal, Puigcerver and Noya-García, 2019), as well as wild-card and approximate or flexible spelling are also supported in the real search systems. However, formal evaluation results for these complex queries still needs fundamental work do define adequate metrics and ecvaluation protocols.

References

Ball, G. R., Srihari, S. N., Srinivasan, H. et al. (2006), Segmentation-based and segmentation-free methods for spotting handwritten arabic words, *in* 'Tenth Int. Workshop on Frontiers in Handwriting Recognition'.

Barrere, K., Toselli, A. H. and Vidal, E. (2019), Line segmentation free probabilistic keyword spotting and indexing, *in* 'Iberian Conference on Pattern Recognition and Image Analysis', Springer, pp. 201–217.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H. and Vidal, E. (2017), Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project, *in* 'Int. Conf. on Document Analysis and Recognition (ICDAR)', Vol. 01, pp. 311–316.

[8] See details in the *"Help and examples"* link of these systems.

Bosch, V., Toselli, A. H. and Vidal, E. (2012), Statistical text line analysis in handwritten documents, *in* 'Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on', IEEE, pp. 201–206.

Cao, H., Bhardwaj, A. and Govindaraju, V. (2009), 'A probabilistic method for keyword retrieval in handwritten document images', *Pattern Recognition* **42**(12), 3374–3382.

Fischer, A., Keller, A., Frinken, V. and Bunke, H. (2012), 'Lexicon-free handwritten word spotting using character HMMs', *Pattern Recognition Letters* **33**(7), 934 – 942. Special Issue on Awards from ICPR 2010.

Frinken, V., Fischer, A., Manmatha, R. and Bunke, H. (2012), 'A Novel Word Spotting Method Based on Recurrent Neural Networks', *IEEE Trans. on Pattern Analysis and Machine Intelligence* **34**(2), 211 –224.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009), 'A Novel Connectionist System for Unconstrained Handwriting Recognition', *IEEE Transaction on Pattern Analysis and Machine Intelligence* **31**(5), 855–868.

Kamel, I. (2010), 'On indexing handwritten text', *Int. Journal of Multimedia and Ubiquitous Engineering* **5**(2).

Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R. and Viorel Popescu, G. (2000), 'A Line-Oriented Approach to Word Spotting in Handwritten Documents', *Pattern Analysis & Applications* **3**, 153–168. 10.1007/s100440070020.

Lang, E., Puigcerver, J., Toselli, A. H. and Vidal, E. (2018), Probabilistic indexing and search for information extraction on handwritten german parish records, *in* '2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)', pp. 44–49.

Manmatha, R., Han, C. and Riseman, E. (1996), Word Spotting: a New Approach to Indexing Handwriting, *in* 'Int. Conference on Computer Vision and Pattern Recognition (ICPR '96)', pp. 631–637.

Manmatha, R. and Rothfeder, J. L. (2005), 'A scale space approach for automatically segmenting words from historical handwritten documents', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(8), 1212–1225.

Manning, C. D., Raghavan, P. and Schtze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Papavassiliou, V., Stafylakis, T., Katsouros, V. and Carayannis, G. (2010), 'Handwritten document image segmentation into text lines and words', *Pattern Recognition* **43**(1), 369–377.

Puigcerver, J. (2018), A Probabilistic Formulation of Keyword Spotting, PhD thesis, Univ. Politècnica de València.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2016), 'Querying out-of-vocabulary words in lexicon-based keyword spotting', *Neural Computing and Applications* pp. 1–10.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2020), Advances in handwritten keyword indexing and search technologies, *in* A. Fischer, M. Liwicki and R. Ingold, eds, 'Handwritten Historical Document Analysis, Recognition, And Retrieval-State Of The Art And Future Trends', Vol. 89, World Scientific, pp. 175–193.

Rath, T. and Manmatha, R. (2007), 'Word spotting for historical documents', *Int. Journal on Document Analysis and Recognition* **9**, 139–152.

Rodríguez-Serrano, J. A. and Perronnin, F. (2009), 'Handwritten word-spotting using hidden Markov models and universal vocabularies', *Pattern Recognition* **42**, 2106–2116.

Romero, V., Toselli, A. H. and Vidal, E. (2012), *Multimodal Interactive Handwritten Text Transcription*, Series in Machine Perception and Artificial Intelligence (MPAI), World Scientific Publishing.

Sánchez, J. A. and Vidal, E. (2021), Handwritten text recognition for the EDT project. Part I: Model training and automatic transcription, *in* M. A. Bermejo et al., ed., 'Proc. of the EDT Alicante workshop', To appear.

Terasawa, K. and Tanaka, Y. (2009), Slit style hog feature for document image word spotting, *in* 'ICDAR-09', pp. 116–120.

Toselli, A. H., Romero, V., Vidal, E. and Sánchez, J. A. (2019), Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, *in* '15th Int. Conf. on Document Analysis and Recognition (ICDAR)'.

Toselli, A. H. and Vidal, E. (2013*a*), Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents, *in* 'Proc. of the Int. Conf. on Document Analysis and Recognition (ICDAR'13)'.

Toselli, A. H. and Vidal, E. (2013*b*), Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents, *in* 'Proc. of the 12th Int. Conference on Document Analysis and Recognition (ICDAR '13)', IEEE Computer Society, Washington, DC, USA, pp. 501–505.

Toselli, A. H., Vidal, E., Puigcerver, J. and Noya-García, E. (2019), 'Probabilistic multi-word spotting in handwritten text images', *Pattern Analysis and Applications* **22**(1), 23–32.

Toselli, A. H., Vidal, E., Romero, V. and Frinken, V. (2016), 'HMM word graph based keyword spotting in handwritten document images', *Information Sciences* **370-371**, 497–518. Information Sciences 370-371 (2016) 497-518.

Vidal, E., Romero, V., Toselli, A. H., Sánchez, J. A., Bosch, V., Quirós, L., Benedí, J. M., Prieto, J. R., Pastor, M., Casacuberta, F., Alonso, C., García, C., Márquez, L. and Orcero, C. (2020), The carabela project and manuscript collection: Large-scale probabilistic indexing and content-based classification, *in* '17th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)', pp. 85–90.

Vinciarelli, A., Bengio, S. and Bunke, H. (2004), 'Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 709–720.

Wshah, S., Kumar, G. and Govindaraju, V. (2012), Script independent word spotting in offline handwritten documents based on hidden markov models, *in* 'Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on', pp. 14–19.

Browsing through sealed historical documents: non-invasive imaging methods for document digitization

D. Stromer[1], M. Seuret[1], J. Schür[2], K. Root[2], I.Ullmann[2], P. Zippert[3], F. Binder[3], S. Funk[4], B. Akstaller[4], L. Dietrich[4], V. Ludwig[4], S. Schreiner[4], M. Schuster[4], D. Haag[4], S. Schmidt[4], T. Michel[4], M. Vossiek[2], T. Hausotte[3], G. Anton[4], A. Maier[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[2] Institute of Microwaves and Photonics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[3] Institute of Manufacturing Metrology (FMT), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[4] Erlangen Centre for Astroparticle Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

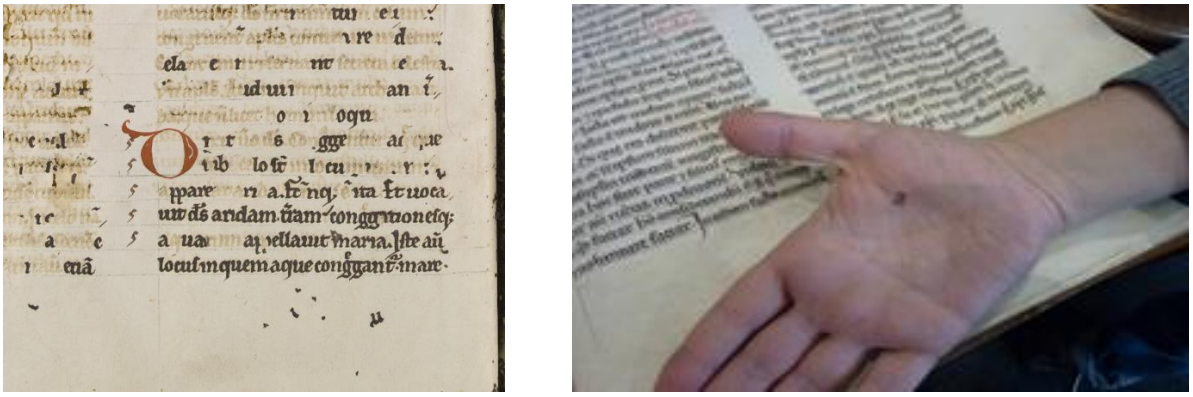
Abstract. *Historical documents are witnesses of history that provide us with valuable information about former times. Many of these relics are too fragile to open them, such that innovative non-invasive imaging techniques can help to reveal hidden contents. In this work, we present our research on Computed Tomography, Phase-contrast and Terahertz imaging. We use image processing methods to visualize the digital data for the naked eye. Our use cases are mainly books, but also Asian bamboo scroll data is shown. As an outlook, our future research will focus on hybrid imaging approaches combined with intelligent image processing. Our research aim is to gain insights, and based on them, provide guidelines for specific documents. Therefore, the space of documents and modalities is presented. We try to utilize advantages and counter disadvantages of certain modalities. Finally, the future of this highly translational research is discussed and possible considerations for potential commercialization are* ***presented.***

1. Introduction and Motivation

Historical documents are witnesses of history that provide us with valuable information about former times. In Germany alone, there are more than 7,770 libraries (as of 2016) with around 218 million visitors annually providing workplaces for around 200,000 people [1]. As a place of educational and cultural mediation, many of these libraries are currently interested in digitizing their existing collections. Innovative technologies are used to for digitization purposes, such as digitally searching for contents in the database.

Books and manuscripts that are in a good condition are partly digitized automatically by using scanning robots. Subsequently, document processing algorithms like OCR are applied to make information findable, assign drawings, or verify writers of books. External influences such as fires can cause massive damage to well-preserved collections at any time and put the documents in a condition that prohibits any contact. The fire at the Herzogin Anna Amalia Bibliothek zu Weimar in 2004 serves as an example. The fire and firefighting efforts severely damaged about 62,000 volumes and left them in a very poor condition, and the majority of these documents had not yet been digitized before the accident [2]. Aging processes can also make it impossible to digitize books with scan robots [3]. A common case in libraries is that pages are stuck in the area of the book fold due to aging; instead of placing such a book into a scan robot, all pages would first have to be separated by experts in a time-consuming manual process.

Furthermore, it is possible that the individual written letters may become detached from the book page, as shown in Fig. 1. The manuscript from Saint-Omer (France) is currently at the Institut de Recherche et d'Histoire des Textes in Paris, France. This document was opened, and after it was observed that the letters were peeling off, it was immediately closed and not processed since then.



(a) (b)

Fig. 1: (a) Page of a historical manuscript currently stored in Paris. After opening, letters have peeled off leaving gaps in the text. (b) A peeled off letter "m". (Source: Institut de Recherche et d'Histoire des Textes, Paris, France).

Since such highly fragile documents can only be digitized by the scan robot method at enormous expense, if at all, research has been done in recent years to find a non-invasive, non-destructive document capture method that would result in a digital three-dimensional volume, allowing a view into the document without opening it manually.

Our research focuses on three well-known individual non-invasive imaging techniques for document digitization. Namely, we work on 3D X-ray Computed Tomography (CT), Phase-contrast Imaging, and 3D Terahertz Imaging. Each of these modalities have their advantages and downsides when acquiring digital representations of an object. Our goal is to combine the advantages and counter disadvantages of individual scans by using a hybrid scanning approach. With this, we want to leverage the optimal non-invasive digitization method for a given document, and finally, provide a scanning guideline for

future research. The scan itself shall then be smartly combined and image processing algorithms will be used to make the sealed information virtually readable for the naked eye. The unique characteristic of our work is the fact that we are the first collaborative community that researches on a multi-modal approach and comparison for individual documents. Therefore, we also filed a patent [4,5]. Furthermore, we also look into damage that could arise from these techniques, e.g., by applying ionizing radiation. Until now, most work only focuses on a specific imaging technique, where there is no comparison to other approaches and no consideration of harming the documents. We also cover the full pipeline, from starting with detailed material analysis, over scanning to image pre- and post-processing. This article is based on experiments described in detail in further publications. The work on Book-CT has been published in [6]. The work on bamboo scrolls in [7]. Both experiments were extended and are described in [8]. This work gives a short review of these articles and extends them by describing our future vision.

2. Documents

Books

The research our group is conducting is mainly based on books. For our experiments, we use self-made books, as shown exemplary in Fig. 2(a). The pages are of handmade paper (Cellulose) and have a mean thickness of 150 µm. The book cover is made from buffalo leather. For writings, we used the following inks: Iron Gall ink ($FeSO_4$), Malachite ink (Cu), Tyrian Purple ink, Buckthorn ink, Indian ink. In this state of our research, we are working on proof-of-concepts with imaging modalities that are normally used for other purposes, such as material testing or medical imaging. Therefore, we have to vary the dimensions of the books for measuring, as the scanner trajectories can mostly just fit small sizes of samples.

(a) (b)

Fig. 2: (a) Exemplary self-made book consisting of 56 handmade paper pages, writings of different inks and a buffalo leather cover. The dimension of the book is approx. 17 cm × 13 cm × 3 cm (length × width × height). This book was used for the Computed Tomography experiments. (b) Bamboo scroll covered in potting soil. The scroll was completely soiled, rolled up, and put in a plastic bag for measurements with the imaging

modality.

Bamboo Scrolls

In addition to books, we also conducted experiments with Asian documents. Bamboo scrolls were widely used before the invention of paper in China during the Han dynasty. In our example, 32 bamboo slips form a complete scroll, where the individual slips are bound together by strings. The slips have a size of 1.2 cm × 15 cm × 0.3 cm (length × width × depth). The wrapped-up scroll has an average diameter of 5.5 cm. Chinese characters and drawings are carved into the surface of the bamboo. The raw bamboo is composed of cellulose, hemicellulose, lignin, ash, and other extractives. Exemplary, we contaminated the scrolls with slightly wet potting soil, as shown in Fig. 2(b). The potting soil is a mix of cellulose and minerals. Here, the minerals have rather high densities where those of cellulose are rather low. To create the worst-case scenario, we pressed the soil onto the recto to fill the air gaps of the carvings. To prevent the scanner from contamination, we put the scrolls into plastic bags that were filled with potting soil and measured the bags with CT.

3. Individual Imaging Techniques and Image Processing Industrial Computed Tomography

Industrial CT has many applications in the field of non-destructive testing. It is one of the few technologies that allows visualizing the internal and external structures of a component as a holistic measurement. In this imaging technique, X-rays pass through an object from different angles and are attenuated by the object. This attenuation depends, among other factors, on the object density and the transmission length within the object. The X-ray intensity attenuation can then be detected and converted into image data [9].

Since historical documents were often written using metallic inks, which differ in attenuation from paper pages, the written letters can be visualized by scanning a document using a CT system (c.f. Fig. 3**¡Error! No se encuentra el origen de la referencia.**). It has already been shown that this technology provides good results for digitalizing ancient books or damaged historical scrolls [10,11]. This is particularly advantageous when historical documents are fragile, and it is no longer possible to open or unroll them without introducing further damage. However, it is not yet possible to simply place any ancient book into the CT system and obtain a digital document of the book content. There are several issues that need to be considered, which is part of our current research.

An important aspect we are investigating is the influence of X-rays on scanned documents. Since CT measurements are based on the effect of ionizing radiation, it must also be considered, whether the X-rays can damage the historical documents during the CT scan. However, this is mainly dependent on the applied radiation energy and this can be reduced by adjusting the measurement settings and applying optimized evaluation algorithms [12].

In addition to scanning books, we also investigated digitization of Asian bamboo scrolls by means of CT. Fig. 4(a) shows the measurement setup, where the soiled scroll was placed on the turntable to acquire the images for the reconstruction. The 3D reconstructed soiled scrolls can be seen in Fig. 4(b), where the writings are visible by the naked eye.

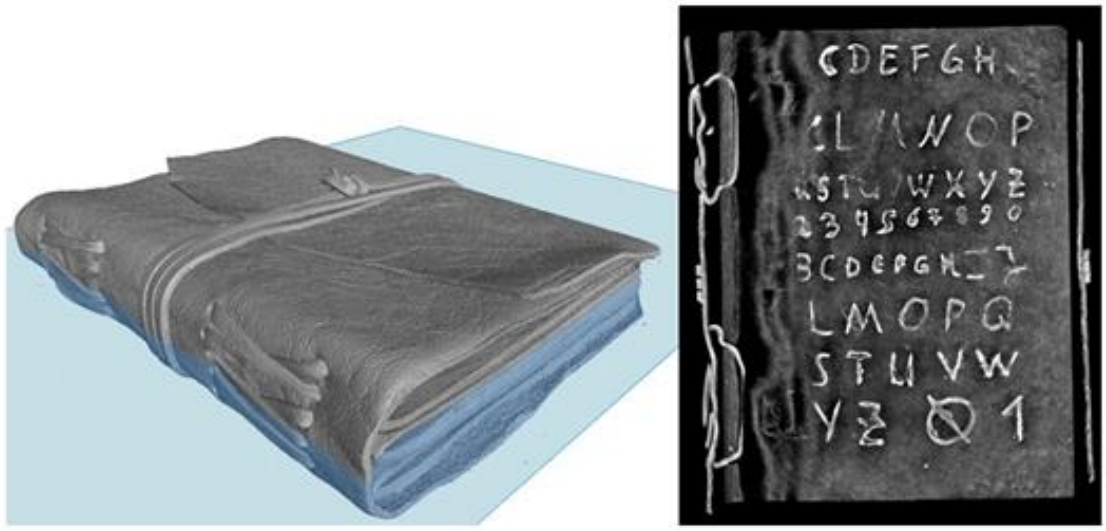
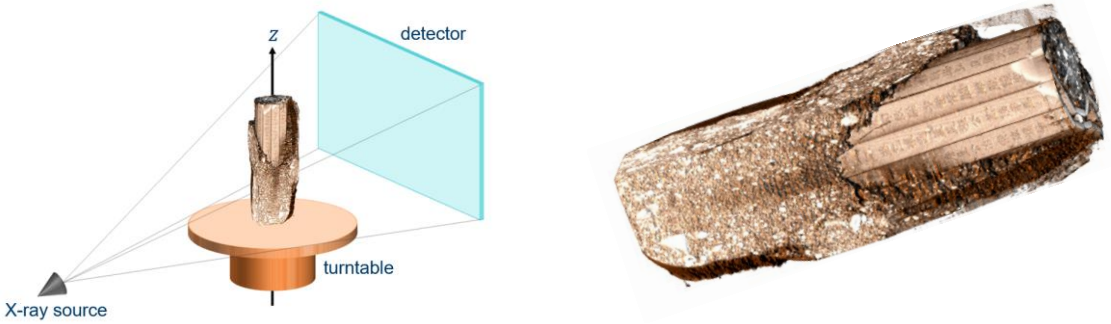


Fig. 3: Visualization of the CT reconstructed volume data of a scanned book, shown in Fig. 2(a). The left image shows the rendered, digital volume. The right image shows a slice through the volume data (blue layer of left image). The letters written with iron gall ink are clearly visible.



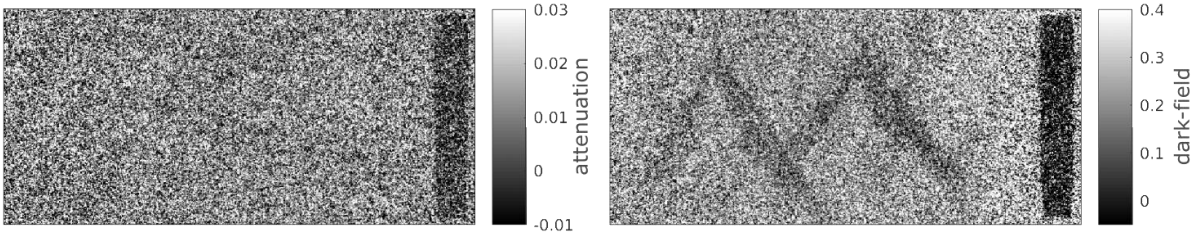
(a) (b)

Fig. 4: (a) The soiled bamboo scroll is positioned on the turntable of the CT scanner for acquiring the images. (b) The reconstructed soiled scroll, rendered for improved visualization. The writings can be seen by the naked eye where the scroll was not covered by soil.

Phase-contrast Imaging

X-ray phase-contrast imaging is an imaging technique which leads to information about both, the object's attenuation and the object's refractive properties [15]. For this purpose,

microstructured gratings are placed into the X-ray beam path and build a Talbot-(Lau) interferometer [16,17]. In common X-ray imaging, the attenuation coefficient μ comprises the image information. By X-ray phase-contrast imaging also the phase shift is accessible. X-ray phase-contrast imaging takes advantage of the so-called Talbot effect. The grating behind the object imprints a phase-shift on the incoming X-ray wave. In certain distances behind the grating, self-images of the grating pattern occur. In such a so-called Talbot pattern, local changes in the X-ray wave front caused by the object are encoded. Besides the known attenuation image, additionally the differential phase and the dark-field image are extracted by analyzing the obtained Talbot pattern. The differential phase image is based on the deflection of X-rays passing a sample. For light elements of similar density, the contrast in a differential phase image is enhanced compared to the common attenuation image [16,18]. This could be an advantage for the investigation of certain combinations of ink and paper types. X-ray phase-contrast imaging already showed its potential with impressing results about the digitalization of carbonized Herculaneum papyrus rolls in Mocella *et al*. and Bukreeva *et al.* [19,20]. In both cases, monochromatic radiation from a synchrotron beamline was used. The dark-field image enables a visualization of scattering structures which are smaller than the resolution of the detection system [21]. Because of the fibrous structure of the paper, there are two structural effects caused by writing: The ink soaks into the paper and the texture of the paper changes by the pressure of the writing tool. Hence, also without any ink remains, it is possible to reveal written symbols by the X-ray dark-field image (Fig. 5). Since the attenuation properties are hardly changed, the respective image does not show the letters. Furthermore, the dark-field image should lead to advantages if the ink contains granular particles. The results were obtained with a laboratory X-ray source using a 30 kVp spectrum and a flat-panel detector of pixel pitch 49.5 μm.



(a) (b)

Fig. 5: Attenuation image (left) and dark-field image (right) of a letter 'M' written without ink. The pressure of the writing tool on the paper leads to a changed texture of the paper which is visible in the dark-field image, but not in the attenuation image

Since X-ray imaging is based on ionizing radiation, which is suspected to cause faster aging of cellulose [22], investigations regarding the applied dose are performed within the project, too. The as-low-as-reasonably-achievable (ALARA) principle is known from the medical field, which aims to obtain all relevant information at the lowest possible dose. To protect the ancient and precious books, an evaluation of the dose is obligatory. Thus, simulation studies are done, where a block of cellulose is placed and irradiated with photons. A deposition spectrum is to be determined within a small volume within the

block. In combination with real measurements, the aim is to gain a measure of the deposited dose and its respective damage in paper in dependence of arbitrary settings of the X-ray tube and resulting spectra.

Terahertz Imaging

Radar imaging is a very well-known approach to use electromagnetic (em-) waves for imaging applications. Satellite based or airborne radar have been and are still used for earth observation or environmental investigations due to the very good propagation conditions in various frequency band with the microwave region and the scattering behavior of the earth surface at these frequencies. In contrast to these remote sensing applications, nowadays short range or near field imaging radars in the millimeter wave frequency range (especially around 80 GHz) are used for security scanners and non-destructive testing. The higher frequency allows imaging with resolutions at mm scales and qualifies this technology for finding thread objects at security checks or voids and defects in materials or products. Imaging and artificial reading of historical documents is not feasible, as the imaging quality is not good enough for highly detailed objects like printed or handwritten documents. Increasing the radar frequency to the THz frequency range (~100 GHz – 10 THz) can overcome the resolution problem and remain the positive propagation conditions as electromagnetic waves in this frequency range are still able to penetrate a large variety of non-conductive (i.e., dielectric) materials among them paper and papyrus. The ability to penetrate a stack of papers is one prerequisite to analyze the written content of a documents without the need to unfold it. The second important aspect is that the ink has to deliver a decent imaging contrast compared to the page material weather by high absorption, high reflectivity of em-waves or by a high enough difference of its index of refraction or complex permittivity respectively. The qualification of THz imaging for the analysis of documents and painting has already been proven [23,24], but not for handwritten documents with ink. The current work described here addresses the question on how suitable the THz frequency range is for this task and what imaging approach (transmission, reflection imaging, SAR-based radar tomography) delivers the most information of the document and may be combined with AI character recognition. In a first experiment different water-based iron gall writing inks (Rohrer & Klingner, black ink 40710 & blue 40711), a resin-based drawing ink (Rohrer & Klingner, black ink 29770) and a generic china ink (black) were analyzed with a commercial imaging system (Rohde & Schwarz QAR) at a frequency around 80 GHz in a reflection setup.

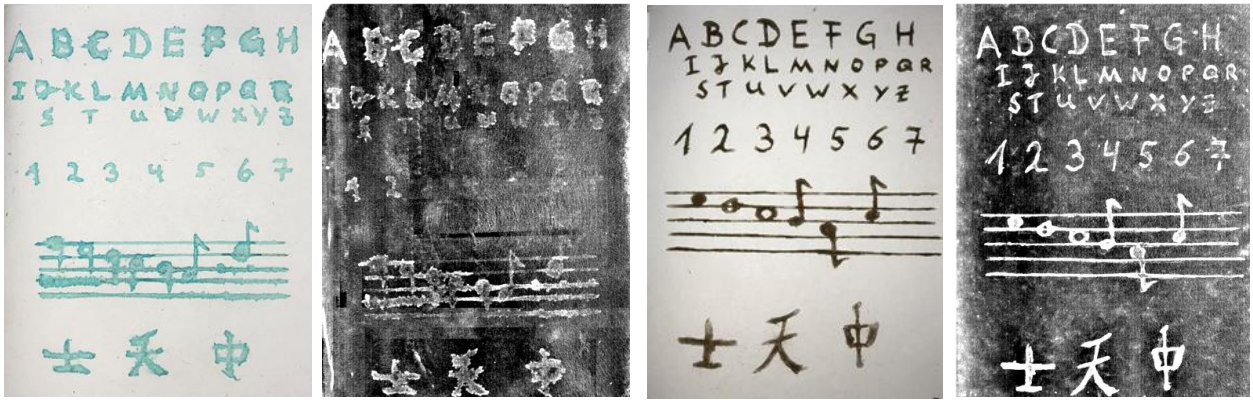
(a) (b)

Fig. 6: (a) Photograph of two sample holders (top row) and radar image (bottom row). (b) Photograph of paper with two letters and corresponding radar image.

In Fig. 6(a), photographs of two sample holders (top row) with the following ink samples are shown: left sample with three inks (iron gall black (top left), iron gall blue (top right) and resin-based black ink (bottom)), right sample with a Chinese inkstone circle (diameter of samples approx. 15 mm). In the radar images below, only the resin-based ink shows a contrast. To evaluate improvements in the imaging quality when using higher frequencies, a sample with two handwritten letters (resin-based black ink) on a single page has been measured in a 220 – 325 GHz lab imaging system [25]. Fig. 6(b) shows a photograph of a paper with two handwritten letters (Letter “L” on recto and “N” on verso). In the radar image both letters can be identified but the image also shows artifacts resulting from undulations of the paper. These preliminary results show that THz imaging of documents is potentially able to artificially read ink handwritings without the need to necessarily have physical or visual access to the specimen. The next steps are experiments with a THz TDS system. The bandwidth of such a system is usually in a region in which the range resolution is sufficiently high to separate single pages in the volume of interest. Advanced signal processing approaches like time domain reflectometry and SAR-base image reconstruction will be implemented and figures of merit will be derived to support the creation of a scanning guideline for future research.

Image Processing

In addition to the imaging methods, the digital data provided by the modalities have to be processed adequately. Details on the image processing algorithms can be found in [7]. Fig. 7 (a,c) shows photographs of sample pages from the books with different metallic inks. Fig. 7 (b,d) show the same page, digitized by CT imaging of the closed book after applying the developed image processing prototype. The pages have been virtually flattened such that the writings are made visible without opening the book manually. The metal particles of the ink generate a contrast that makes it possible to distinguish ink from the cellulose-based pages.



(a) (b) (c) (d)

Fig. 7: (a) Original page of a sample book with writings of Malachite ink. (b) 3D CT reconstructed and image processed page. (c) Original page of a sample book with writings of Iron Gall ink. (d) 3D CT reconstructed and image processed page. The images have been extracted by an algorithm and the writings are clearly visible and readable.

For the bamboo scrolls, we developed a different processing pipeline. First, the scanned scroll is getting virtually cleaned from the soil and then unwrapped by a second algorithm. Finally, the unwrapped bamboo elements get sampled at the dense surface such that the carvings are made readable. The results are shown in Fig. 8, where the carvings are made digitally visible for the naked eye.

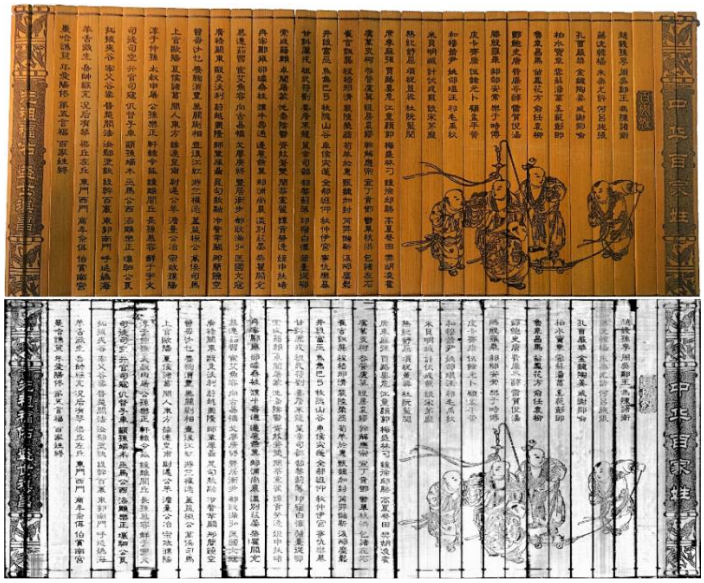


Fig. 8: (Top) Photograph of unwrapped scroll. (Bottom) Virtually cleaned and unwrapped scroll. The carvings are made readable by erasing the soil from the volume and unwrapping the bamboo elements.

4. Hybrid Imaging

Until now, most investigations are performed by using a single modality for a certain document. To the best of our knowledge, there are no publications on hybrid/multi-modality non-invasive imaging experiments performed on the same document so far. Our research focuses on combining the advantages of the presented imaging techniques to counter their disadvantages, and ultimately, use image processing methods such as registration or segmentation to provide a visual output of the digitally acquired data.

Therefore, we developed a sketch, shown in Fig. 9, demonstrating the space of document digitization and the attributes that play important roles for the success of the methods.

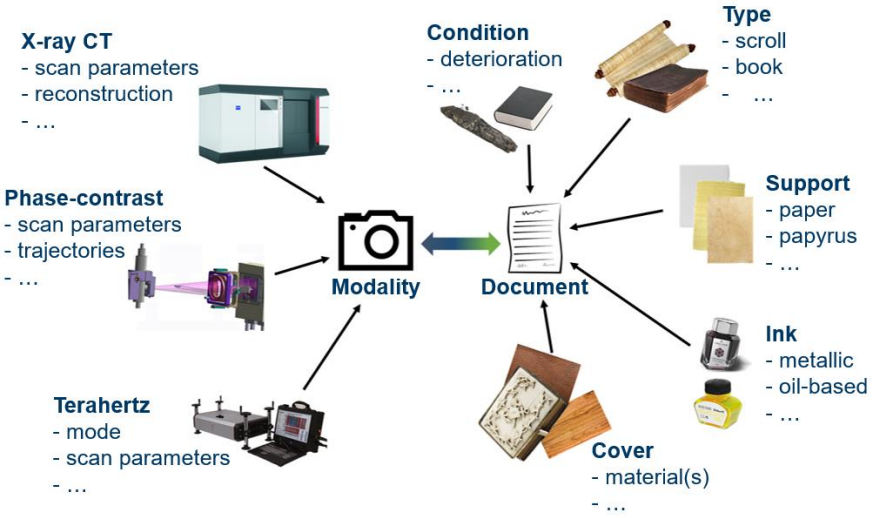


Fig. 9: The space of digitizing documents can be divided into two main fields. The document itself has different types of attributes such as condition, type, cover, ink, etc. On the other hand, different modalities exist such as CT, Terahertz or Phase-contrast.

In our project, we are working on guidelines that provide information on measurement techniques with certain documents. The (open source) document database shall grow by time, as well as the modality and measurement parameter information. The outcome of our research will be a standard operating procedure (SOP) cookbook for a given document, delivering automatically measurement parameters and other considerations that have to be taken into account.

5. Discussion and Outlook

For the research we conduct, the problem space is manifold. The project is highly translational as there is a need for lots of expertise of very different fields such as humanities, electrical engineering, material science, physics, chemistry or computer science. Only with collaboration between these fields, valid outcomes will be generated.

Next, most documents that are hardly digitizable are unique and have very specific conditions. Creating scalable solutions for digitization ending up in concrete business

models is a very challenging task, as also the scanners are normally made for imaging at certain labs in closed environments. In contrast, carrying the fragile documents to measurement center is sometimes not feasible as insurance and transportation costs will be very high. Therefore, when thinking of commercializing the techniques we present, one would have to identify valid use cases that justify sustainable business in terms of monetization. Also, interdisciplinary research and development will be key.

If in the future, more and more research in this relatively new field would be conducted and methods evolve, there will be seen parallels to the medical field when it comes to leveraging AI methods for gaining insights into the digital data. Learning from that field when it comes to data privacy regulations or ethics can be leveraged and extended to this use case for future products and business models. When it comes to artificial intelligence, federated learning would be a great principle to gain deeper insights into the data to build connections and train machine learning algorithms.

6. Conclusion

In this article, we present our current research on using different non-invasive imaging modalities such as CT, Phase-contrast and 3D Terahertz for gaining visible insights into closed documents. To reveal the hidden content, we use image processing algorithms subsequently to the image acquisition. In future, we aim to combine these methods to utilize the advantages of the different techniques and provide guidelines for digitization.

Acknowledgements

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the founding of the research project 433501541 and also for the financial support of the acquisition of the CT system Zeiss Metrotom 1500 through Grant No. 324672600.

References

[1] Heugen-Ecker Gabriele. “DBS – Deutsche Bibliotheksstatistik 2016.” [Online; accessed 22-May-2019]. 2016. https://wiki1.hbz-nrw.de/download/attachments/99811333/dbs_gesamt_dt_2016.pdf?version=1&modificationDate=1508326213410.

[2] Knoche Michael. “The Herzogin Anna Amalia library after the fire.” IFLA journal 31, no. 1 (2005): 90-92.

[3] Coyle, Karen. “Mass digitization of books.” The Journal of Academic Librarianship 32, no. 6 (2006): 641-645.

[4] Schoen Tobias, Holub Wolfgang, Stromer Daniel, Maier Andreas, Anton Gisela, Thilo Michel, Vossiek Martin, and Schuer Jan. “System for analyzing a document and corresponding method.” WO2019038403. 2018.

[5] Schoen Tobias, Holub Wolfgang, Stromer Daniel, Maier Andreas, Anton Gisela, Thilo Michel, Vossiek Martin, and Schuer Jan. “System for analyzing a document and corresponding method.” U.S. Patent 11,057,536, issued July 6, 2021.

[6] Stromer Daniel, Christlein Vincent, Martindale Christine, Zippert Patrick, Haltenberger Eric, Hausotte Tino, and Maier Andreas. “Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources.” *Scientific reports* 8, no. 1 (2018): 1-10.

[7] Stromer Daniel, Christlein Vincent, Huang Xiaolin, Zippert Patrick, Hausotte Tino, and Maier Andreas. “Virtual cleaning and unwrapping of non-invasively digitized soiled bamboo scrolls.” *Scientific reports* 9, no. 1 (2019): 1-10.

[8] Stromer Daniel. “Non-invasive imaging methods for digital humanities, medicine, and quality assessment.” PhD diss., Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2019.

[9] Carmignato Simone, Dewulf Wim, and Leach Richard, eds. “Industrial X-ray computed tomography.” Cham: Springer International Publishing, 2018.

[10] Albertin Fauzia, Patera Alessandra, Jerjen Iwan, Hartmann Stefan, Peccenini Eva, Kaplan Frédéric, Stampanoni Marco, Kaufmann Rolf, and Margaritondo Giorgio. “Virtual reading of a large ancient handwritten science book.” *Microchemical Journal* 125 (2016): 185-189.

[11] Rosin Paul L., Lai Yu-Kun, Liu Chang, Davis Graham R., Mills David, Tuson Gary, and Russell Yuki. “Virtual recovery of content from X-ray micro-tomography scans of damaged historic scrolls.” *Scientific reports* 8, no. 1 (2018): 1-10.

[12] Stromer Daniel, Christlein Vincent, Huang Yixing, Zippert Patrick, Helmecke Eric, Hausotte Tino, and Maier Andreas. “Dose reduction for historical books digitization by 3-D X-ray CT.” In *of Applied Sciences Upper Austria, U.(ed.) Proceedings of 8th Conference on Industrial Computed Tomography*, pp. 1-2. 2018.

[13] ASTM, ASTM E. “2597: Standard Practice for Manufacturing Characterization of Digital Detector Arrays.” (2007).

[14] Lifton Joseph J., Malcolm Andrew A., McBride John W., and Cross Kevin J.. “The application of voxel size correction in x-ray computed tomography for dimensional metrology.” In *Singapore international NDT conference & exhibition*, pp. 19-20. 2013.

[15] Bonse Ulrich, and Hart Michael. “An X-ray interferometer.” *Applied Physics Letters* 6, no. 8 (1965): 155-156.

[16] Momose Atsushi, Kawamoto Shinya, Koyama Ichiro, Hamaishi Yoshitaka, Takai Kengo, and Suzuki Yoshio. “Demonstration of X-ray Talbot interferometry.” *Japanese journal of applied physics* 42, no. 7B (2003): L866.

[17] Pfeiffer Franz, Weitkamp Timm, Bunk Oliver, and David Christian. “Phase retrieval and differential phase-contrast imaging with low-brilliance X-ray sources.” *Nature physics* 2, no. 4 (2006): 258-261.

[18] Fitzgerald Richard. “Phase-sensitive X-ray imaging.” *Physics today* 53 (2000): 23-28.

[19] Mocella Vito, Brun Emmanuel, Ferrero Claudio, and Delattre Daniel. “Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging.” *Nature communications* 6, no. 1 (2015): 1-6.

[20] Bukreeva Inna, Mittone Alberto, Bravin Alberto, Festa Giulia, Alessandrelli Michele, Coan Paola, Formoso Vincenzo *et al*. “Virtual unrolling and deciphering of Herculaneum papyri by X-ray phase-contrast tomography.” *Scientific reports* 6, no. 1 (2016): 1-7.

[21] Yashiro Wataru, Terui Yoshiharu, Kawabata Keisuke, and Momose Atsushi. “On the origin of visibility contrast in x-ray Talbot interferometry.” *Optics express* 18, no. 16 (2010): 16890-16901.

[22] Charlesby Arthur. “The degradation of cellulose by ionizing radiation.” *Journal of Polymer Science* 15, no. 79 (1955): 263-270.

[23] Jackson J. Bianca, Bowen John, Walker Gillian, Labaune Julien, Mourou Gerard, Menu Michel, and Fukunaga Kaori. “A survey of terahertz applications in cultural heritage conservation science.” *IEEE Transactions on Terahertz Science and Technology* 1, no. 1 (2011): 220-231.

[24] Fukunaga Kaori. “Thz technology applied to cultural heritage in practice.” *Springer*, 2016.

[25] Ullmann Ingrid. “Contactless Inspection of Handwritten Documents with Terahertz Imaging.” Submitted for the *European Microwave Week 2021*, London (2021).

LINK-LIVES: BUILDING HISTORICAL BIG DATA FROM ARCHIVAL RECORDS FOR USE BY RESEARCHERS AND THE DANISH PUBLIC

Bárbara Ana Revuelta-Eugercios

Project senior researcher at the *Danish National Archives* / Research Associate Professor at the *University of Copenhagen*
bre@sa.dk, +45 41 77 74 88

Abstract. *The Link-Lives project, which is a cross-disciplinary research project, will take the difficult and time-consuming task of combining information from diverse archival sources relating to any given person, to build life-courses and family relations from 1787 to the present and make them freely and easily available. This will, on the one hand, expand the scope of registry-based research from decades to centuries and open up new avenues for intergenerational research in the health and social sciences and, on the other hand, ease the access to some of Denmark's digital treasures to the average citizen. It is a collaboration of the Danish National Archives, the Copenhagen City Archives and the University of Copenhagen. It is funded through two grants by the Innovation Fund Denmark, the Carlsberg Foundation and two small grants by the Ministry of Culture.*

1. Introduction: two user-groups underserved by archives in Denmark

There are currently two user groups still underserved by the Danish archives. On the one hand, the community of researchers from the Humanities, the Social Sciences, and even the health sciences, who are interested in the lifecourse and multigenerational mechanisms that affect the biological and social lives of humans. On the other hand, the large pool of citizens who would like to get started in family history but do not have the time or the competences to do it on their own.

Let me illustrate the first case with the story of fictional Hannah, a PhD student in Humanities or Social sciences at a Danish university, who is interested in the new field of research that focuses on analyzing the intergenerational mechanisms explaining social and biological life. She has read, for example, research using 400 years of Canadian historical parish records that showed that there was a selective advantage to moderate fertility in the frontier population. The paper showed that "while high fecundity was associated with a larger number of children, perhaps paradoxically, moderate fecundity maximized the number of descendants after several generations" (Galor and Klemp 2019). She has also read about how the adverse health consequences of a birth out of wedlock in Sweden in the 1920s were not restricted to the child himself but could also be felt in their offspring and their offspring's offspring, the later born in the late 20th century (Modin, Koupil, and Vaguero 2006). Also, how US researchers showed that being a child in the US in the

1940s and living in areas with a variety of ethnic groups was related to particular political affiliations 7 decades later (Brown et al. 2021).

Impressed with the possibilities of this type of research, she wonders: “Could these studies be done in Denmark?” She knows that the Central Person Registry, which was established in Denmark in 1968, has very high quality registry data. This registry connects all personal information of residents in Denmark through a unique personal identification number, which allows all types of registries to be combined. The challenge is how to reconstruct the lives of generations before 1968. As soon as she starts googling historical sources, the fantastic news for Hannah is that the Danish archives, national, municipal and local hold a wealth of treasures, that could indeed allow reconstructing lifecourses and generations all the way back to the first census of 1787 – for some areas, even before in time, to the first parish registries that started in the 16th century. The even better news is that there are millions of images of these sources already scanned and photographed, many of them freely available, and also of millions and millions of transcribed records.

Indeed, there are individual and household data from the census sheets and the parish records are fully available as digital facsimiles, together with full transcriptions of more than 10 censuses, in the Danish Demographic Databases, the result of a crowdsourcing project that started in 1992 (Clausen 2015). Other sources of images and transcribed data are, for instance, the municipal archives, which have also run their own digitation and crowdsourcing projects making other local sources available. Additionally, genealogical companies, e.g., Ancestry, have also created new images and transcriptions, for instance, of the parish records, which have just made freely available in Denmark. Many research projects have also created historical data registries, especially for health, that contain precious information on the early 20th century. And, last but not least, there is a host of small databases held by genealogical associations, private citizens and other actors that also have digitized and transcribed enormous amounts of data from the archives.

However, while the resources are there and reconstructing lifecourses and generations back in time is entirely possible, the task is way out of Hannah’s reach and her project’s or of any other single researcher or research project for that matter. There is simply too much data in very different formats, qualities, etc., which makes it impossible for her to fully take advantage of this digital treasure trove because, to all intents and purposes, it is constituted of isolated digital islands of information. Hannah may well decide to try to acquire Canadian, Swedish or American data for her research instead.

The second case I want to illustrate is that of Hans, a middle-aged blue-collar worker who has always been interested in family history. After the birth of his grandson, he finally decides to make a family genealogy and he googles “how to get started in family history”. The first thing that shows up is the company MyHeritage, well established in Denmark, and he gets very excited as it all seem very easy. Soon, however, he hits a paywall, abandons My Heritage – as he is not yet willing to pay for it, - and googles some more and starts finding some of the resources that Hannah found: the National archives, the municipal archives, the genealogy pages, forums, Facebook… What is the difference between pages, why are there so many versions of the censuses, … Where

does he start? Every link sends him to more and more opportunities and more datasets and images of descriptions or pages… So, finally, it becomes too much for him, there is too much information, too much to figure out, and he does not have the time. He many decide that maybe when he retires, he will have a look at it, when he will have the time.

2. The *Link-Lives* project

In short, the *Link-Lives* projects aims at solving the needs of these two types of users and in doing that, also at making data available in a format that may attract even more type of users: the general public, educators and students, and other archives and cultural heritage institutions. Link-Lives is a cross-disciplinary research project that will take the difficult and time-consuming task of combining information from diverse archival sources relating to any given person, to build life-courses and family relations from 1787 to the present and make them freely and easily available. This will expand, on the one hand, the scope of registry-based research from decades to centuries, and open up new avenues for intergenerational research in the health and social sciences, and, on the other hand, will ease the access to some of Denmark's digital treasures to the average citizen. There is a large demand of this type of datasets placed on the institutions that have been able to create them. For instance, more than a thousand articles have been written on data from the databases held at CEDAR, which covers some areas of Northern Sweden. More and more researchers are trying to find multigenerational and genealogical data, (the later, even if flawed) to conduct this type of research (Song and Campbell 2017; Ruggles 2014; Kaplanis et al. 2018), and new projects focusing on intergenerational aspects are being funded (genpop n.d.).

In order to create these lifecourses we combine machine learning, historical research, bioinformatics and citizen involvement to transform Danish archival sources into multigenerational big data. This endeavor is only possible through the cooperation of the Danish National Archives (*Rigsarkivet*), Copenhagen City Archives (*Københavns Stadsarkiv*) and the University of Copenhagen and the funding obtained through two generous large grants by the Innovation Fund Denmark and the Carlsberg Foundation and two small research grants by the Ministry of Culture. The timeframe for the project is 2019-2024.

The process of converting archival material into big historical data will be done and stored into what we call *Link-Lives Links*, a central data infrastructure hosted at the National Archives, which will be disseminated to users through two services: *Link-Lives Science*, a service to researchers to facilitate research, and *Citizen*, a public webpage, where everyone will be able to search and explore the links and data created by the project – i.e., data not protected by the GDPR and its Danish implementation,- which in effect means that of individuals born before 1901.

In order to avoid that the result of the project becomes a data cemetery, the setup will make the continuous integration of new data possible, even after the project's end. These will come either from additional sources transcribed by the ongoing crowdsourcing projects at the Danish archives, or from agreements with other genealogical societies. They could also come from integration of new archival records covering different dimensions of individual lives transcribed by research projects, as

paid-for service, that will be part of the income-funded activities at the National Archives, where researchers could pay for transcription or linking of their collections to Link-Lives. Thus, the end result is not a dataset but an exponentially growing infrastructure able to incorporate historical data about individuals in a dynamic manner.

The project has two phases, the period 2019-2022, focusing on publicly available data not protected by the GDPR until 1901, followed by the integration of GDPR protected data in 2022-2024. In phase 1 we develop a proof of concept, i.e., that this type of data integration is possible, by linking three sources that have been created by three different actors, two archives and a genealogy company. We are currently linking the 10 fully transcribed censuses held at the Danish National Archives, from 1787 to 1901. These contain information on every person in the country in each census year, comprising more than 10M records, that have been obtained through crowdsourcing. As I have mentioned, this has been made possible by volunteers who have been transcribing and scanning for more than 20 years (Clausen 2015; 'Dansk Demografisk Database' n.d.). We link censuses to each other, but also link them to the parish records that have been indexed by the genealogical company Ancestry covering the period 1812 to 1911 (Van Zeeland and Gronemann 2019). They include information on baptisms, marriages and burials, and contain more than 22M records. We also link to the Copenhagen City Funeral records, which includes all burials in the city between 1860 and 1940, including the person's cause of death, which are being transcribed also as a part of a crowdsourcing project at the Copenhagen City Archives (Van Zeeland and Gronemann 2019). Starting in 2022, we will extend our coverage to the early 20th century, through additional censuses, and connect to the modern CPR registry.

At the end of our current funding, we hope to secure additional funding that will allow us to extend the collection with richer sources from the same two archives but also those of other cultural heritage institutions. This is possible because there is an enormous amount of transcribed archival records that are already waiting to be included. All these ensures the project's growth path in three ways. First, through the contributions made available by volunteers who have scanned and photographed millions of images and also transcribed and indexed many millions of records. The value of these contributions amounts to millions of Danish kroner. Second, through collaborations with private companies. And third, through the increasing amount of research projects whose data could be included. For instance, we have a collaboration with researchers at the University of Southern Denmark to link a new dataset on biographies of students who graduated high school during the 19th century to the main Link-Lives. Also, there are already plenty of transcribed projects, as the Cause of death Register, which contains information on all death certificates for the country starting in 1942 and extending into the period cover by modern registration (Juel and Helweg-Larsen 1999). Moreover there are recently funded projects also with exciting prospects. For instance, there is a new project hosted at the National Archives, the *Multigenerational Registry* (Novo Nordisk Fonden n.d.), which is looking into creating even more records from parish registries, focusing on the transcription of the parish records of Denmark from 1920 to 1968 through the development of new text recognition technologies.

The main institutions behind the project are the National Archives, Copenhagen City Archives and two departments at the University of Copenhagen. The team we have

assembled to carry out of project reflects the wide array of expertise needed to convert historical sources, which start their life as analogue pieces of paper, and ensure multiple transformations through scanning, transcription, standardization and linking until they become historical big data. We have a combination of archivists, historians and historical demographers in close collaboration with data scientists and biostatisticians that ensure that we develop the best methods, those that fully incorporate an accurate knowledge of historical sources, lives, societies and reflect the latest trends on state-of-the-art entity resolution. We also ensure that archivists and historians can engage and disseminate to students, family historians and the public at large.

Moreover, to ensure that our work aligns and builds on the current state of the art in linking, we collaborate with research teams at foreign universities with large experience in linking historical records and carrying out intergenerational research and developing new methods, in Sweden, Norway, Scotland and the Netherlands. Our colleagues at the University of Umeå (Edvinsson and Engberg 2020) and Tromsø (Thorvaldsen, Andersen, and Sommerseth 2015) have been engaged in creating historical demographic databases from historical records from the late 1980s, so they have ample experience in treating and dealing with Nordic material, which tends to share many similarities. The new project *Digitizing Scotland* at the University of Edinburgh is, on the other side, another relative newcomer which is transcribing and linking all vital registration in Scotland for the period 1850-1950, leaning heavily on developing new methods to deal with millions of records (Akgün et al. 2020). The group from the Radboud University Nijmegen have experience developing and working on a variety of historical datasets from Dutch historical records. (Mandemakers and Kok 2020).

In the following sections I describe with more detail how we are linking data and how we are planning to deliver it through our services: *Science* and *Citizen*.

3. *Link-Lives Links*: creating links and lifecourses

Given that the term “linked data” has different interpretations in data management and dissemination, let me briefly explain what we mean when we talk about linking. We are not talking about the Semantic Web in this project but on the field of entity resolution (Christen 2012). A “link” for us is the relationship between two records containing personal information that come from two historical different sources, like two census records from two years that we think belong to the same person. By chaining links from different sources, we can create lifecourses that give a reconstructed image about the events that individuals went through in their lives and, from the contextual information in them, reconstruct also family and kinship relations and reconstruct generations.

While on the outset the process seems very similar to that of genealogy and, in a sense, it is, the way we decide if something is a link is different. Where they cross-check different sources from different types of media following a single individual, we need to implement linking methods that can be scaled up to millions of records, which, for now, leaves us at the level of pair-wise linking, i.e., between records in two sources, which is the standard in the literature in entity resolution now. We use the expression “the most probable link” to acknowledge the fact that we can never know if a link was right or no, as there is no way to go back and check, there is no real “ground truth”, as it is called in

machine learning. Instead, we develop different methodologies that allow us to arrive to our best estimate of whether any two records are a link. And we develop those methods with maximum focus on transparency, reliability and reproducibility, ensuring each link as metadata, so anyone can assess, reproduce or challenge our methods.

In this project we implement three methods of linking. First, we create sets of linked records through manual linking. That means that a human (a domain expert/historian) takes a decision on whether the information from two records fulfills the conditions to be considered a link. We have created a software, *Assisted Linking Application* (ALA), that enables computer-assisted linking and developed a set of protocols and guidelines to ensure systematic data creation. We use two independent linkers whose disagreements are afterwards resolved by an arbiter. As of August 2021, we have created more than 35.000 records for different types of sources, year ranges and areas. The construction of this data is a cornerstone of our work because, in the absence of true ground truth, it is what allows us to have a best estimate of what a domain expert thinks is a link. And this data is then key to test and train automatic models. Our data shows that humans can find up to 80% of cases in most instances, but there is a large variation between geographies, chronologies, sources and users.

The second type of method is a set of rule-based algorithms where a historian and a data scientists program a set of rules that can be implemented across the whole dataset. These have been widely spread in many projects as they are relatively easy to implement (Ruggles 2002; Ruggles, Fitch, and Roberts 2018; Thorvaldsen, Andersen, and Sommerseth 2015; Fu et al. 2014). A simple rule could be that two records need to have a close enough name, place of birth and age to be a link. Of course, the devil is in the details, e.g., how close do they need to be to be considered “close”? Comparing the model results with our domain-expert created data, models capture around 70% of what human domain experts can and the whole linkage program can be run in 2-3 hours in our high-performing environment, while it takes around 3-4 person hours to fully link, compare and resolve 100 records of domain expert data.

The third type of models is comprised of other automatic methods and machine learning approaches. In the simplest of terms, in machine learning, we take the small set of records created by our domain experts (called “training data”) and feed them to a model that, then, figures out from that data what is a link and what is not a link and that can produce prediction, also for the whole dataset. And if we save part of the linked data and do not use it completely to train the model, we can then use it to test how the model compares to our domain experts. We have recently gathered enough and varied-enough training data and are in the process of testing different implementations. We are implementing methods already in use in the literature in order to benchmark and compare them, including the Expectation-Maximization approach (Abramitzky, Mill, and Pérez 2019), support-vector machine (Ruggles et al., n.d.; 2011; Antonie et al. 2014) but also testing other models, e.g., random forest and variation recurrent neural networks. Their results are very encouraging. Each of these methods have their pros and cons, so they can be used for different purposes by our users, depending on their interests.

However, what it is clear is that the linked intergenerational data researchers like our user Hannah would get from Link-Lives will reflect the historical reality where it

originated, but it will have travelled a very long way and experienced substantial transformations that need to be considered. And this is also central part of *Link-Lives*' mission, to ensure that we document and highlight each step of these explicit or implicit human, computing or human and computing decisions that is responsible for the final data: reality was captured because of government decisions; the registration was implemented by statistical offices and individual agents; individuals may have different levels of willingness to accurately report about their lives; archives may have had chronologically and geographically different preservation policies; archives may also have radically different digitation policies; the aims and initiators of crowdsourcing projects affect the design of what and how information was captured, the volunteers may have different levels of competences and willingness to follow the rules; the *Link-Lives* team has taken many decisions to standardize, process and link data as well as dissemination formats.

This complex lifecourse does not mean that the data is not usable or high quality but that, in order to obtain the highest quality research, we need to make available sufficient documentation capturing these different steps, including metadata. Our aim is that our users can always distinguish between our different levels of interpretation and choose what fits them better.

4. *Link-Lives Citizen* and *Science*: delivering data in the format that users need it

Given our understanding that our users will have different interests, profiles, and competences we have designed a dissemination strategy that take these into account.

LINK-LIVES CITIZEN

To cater for the genealogy and volunteer community, we have designed a search function in our webpage where anyone can freely access all our created links and lifecourses. The data from Links that can be made available for the public because it is not protected by the data privacy legislation, which in its Danish implementation protects individual data up to 10 years after their death. We are in the final stages of the construction of the front and backend and we expect to launch by early 2022. As of now, there is only information about the project in our webpage, linklives.dk, but soon it will be possible to do simple and advanced searches on individuals and be able to retrieve the results generated by our different approaches. After searching, users will be able to scroll through both our original sources and re-created lifecourses, which they will be able to explore. The main difference from our page from other types of similar genealogy resources is that we do explicitly present the users with the methods employed for generating every single link in the form of easily accessible metadata. It is very important for us to show that we do not aim to declare who is someone's great-grandmother or attest that these are "real" lifecourses, but just provide different options for users that make it easier for them to find what they are looking for.

A second difference, important for us, is the inclusion of a feedback function that will allow us to both deal with feedback in a structured manner and gather additional data that we could use to further refine our methods. Users will need to be logged in and provide feedback clicking some boxes, answering whether the link is correct and why

they think it is correct (from the sources already present in Link-Lives, from other sources not yet available for us, or from their own research). These manually verified-links from volunteers and interested family historians will create data very different from our domain-expert, guidelines-constrained linked data. However, we believe that they could help us gather data on the multiple-source-checking way of linking employed by genealogists and, when collected with the right metadata and adequately aggregated, can help us understand more on how pairwise linking compares to multiple-source linking and how to improve our models.

In all the process from design to implementation, we have included several rounds of user-testing, not only with family historians and volunteers, but also with researchers and other general public groups, to ensure that we incorporate user feedback to maximize both functionality and ease.

LINK-LIVES SCIENCE

Link-Lives Citizen will serve as a window for researchers to get a first glimpse of the data but they will actually get access through what we call *Link-Lives Science*, where researchers will be able to download all, fully available, data files for sources, links, lifecourses, metadata and documentation. This service will be hosted at the National Archives and will be operational in February 2022. It will provide all data as a simple service, and any researcher with some programing competences will be able to work with it. It will also include annual releases of new linked data until 2024. However, we hope to secure additional funding to develop an easier interface so that also less programming-savvy researchers, students and the public can also engage with the data in different ways. When we include data protected by GDPR protections and its Danish implementation, access will require the same type of permits and secure access as any other data held and requested at the National Archives.

As part of the development of *Science*, the project itself, and with collaborators, is engaging in research with the data as we develop it. This is a way to ensure that the data is tested as it is developed and that the new insights gathered by the research that we perform provides new knowledge about the source, the data or Danish history that can itself be incorporated into the development. The research that is being developed by us and our collaborators touches a variety of disciplines, from history, historical demography, history of medicine, economic history, onomastics, data science, bioinformatics, archival studies… thus, while this project may look like an infrastructure project, its construction is driven by research in historical methods and other disciplines, where most of the team members are engaged in research in one of the 10 articles that we currently have ongoing: I am a senior researcher myself and two professors, two senior researchers, two postdocs and three PhD students,

5. Perspectives: value and beneficiaries

Overall, we believe that there are main two beneficiaries of the projects are; first, the Danish general public/archives, but also the international research community, who will be able to access new data and will open up unprecedented avenues of research. It will be possible not only to do historical research but also to expand the possibilities in

sociology, political science, economics, health sciences and data science. All of this will, at the same time, hopefully attract researchers and research funding to Denmark.

Second, the Danish community of family historians and the general public, who will be able to access freely all that public data in a new form. This will be a way of giving back to the community of volunteers who have transcribed the data through crowdsourcing. We expect that the page will become for many a first “go-to” place for getting started in family history, facilitating the way into the rich ecosystem of resources for genealogy.

Third, for the archives partnering in the project, creating this tool serves as a new way of exploring disseminating their collection and engaging their users. The project provides an opportunity to experiment with a new role of data creation, substantially different from the traditional roles of preservation and dissemination of data. For example, for the National Archives, *Link-Lives* underpins one of the new strategic directions focused on making data available in new ways. For Copenhagen City Archives, *Link-Lives* ensures that the volunteer-research loop mediated by the archive is closed: the engagement of their users in their crowdsourcing projects can be fruitfully used for research, which is later made available for them in new ways, that can lead to new ways of engagement.

Moreover, although it is clear that it is beyond the scope of the project as it is right now, the type of structure we propose opens the door for including other types of digital treasures from other GLAMs (Galleries, Libraries, Archive and Museums) for research or wide-public interest: new archival collections from crowdsourcing in municipal archives, both in the form of traditional registries or in the growing number of letters and other natural-language sources that are becoming available, artists’ lifecourses could be connected to their artworks in Danish museums, writers with their publication at the Royal Library and even local personalities to their contributions or artifacts in local museums or archives.

Finally, while there are projects in the US and Europe carrying out some of the elements that are included in this project, i.e., research projects linking large-scale data, archives engaging in transcription, collaboration with genealogical companies, etc., we think that this project has two differentiating features that make it very strong: first, while we in Denmark are relative newcomers to the business of creating historical databases, which has a very long tradition other university departments in Europe, as our colleagues in Norway and Sweden, the enormous wealth of pre-existing data and the new technological developments, have allowed us to put together a project that basically has progressed from nothing to a full nation-wide population database for the 19th century in three years, becoming/arriving among the first countries in the world to be able to do that. Second, our large-scale established collaboration has allowed us to build a business model that tries to reach different types of users and engage into synergies that, by design, will contribute to propel it further into the future beyond the end of our current funding.

6. References

[1] Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2019. ‘Linking Individuals across Historical Sources: A Fully Automated Approach*’. Historical Methods: A Journal of

Quantitative and Interdisciplinary History, 1–18. https://doi.org/10.1080/01615440.2018.1543034.
[2] Akgün, Özgür, Alan Dearle, Graham Kirby, Eilidh Garrett, Tom Dalton, Peter Christen, Chris Dibben, and Lee Williamson. 2020. 'Linking Scottish Vital Event Records Using Family Groups'. Historical Methods: A Journal of Quantitative and Interdisciplinary History 53 (2): 130–46. https://doi.org/10.1080/01615440.2019.1571466.
[3] Antonie, Luiza, Kris Inwood, Daniel J. Lizotte, and J. Andrew Ross. 2014. 'Tracking People over Time in 19th Century Canada for Longitudinal Analysis'. Machine Learning; Dordrecht 95 (1): 129–46. http://dx.doi.org.ep.fjernadgang.kb.dk/10.1007/s10994-013-5421-0.
[4] Brown, Jacob R., Ryan D. Enos, James Feigenbaum, and Soumyajit Mazumder. 2021. 'Childhood Cross-Ethnic Exposure Predicts Political Behavior Seven Decades Later: Evidence from Linked Administrative Data'. *Science Advances* 7 (24): eabe8432. https://doi.org/10.1126/sciadv.abe8432.
[5] Christen, Peter. 2012. Data Matching, Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. (online): Springer.
[6] Clausen, Nanna Floor. 2015. 'The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data'. In Population Reconstruction, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 3–22. (online): Springer.
[7] 'Dansk Demografisk Database'. n.d. Accessed 22 December 2017. http://www.ddd.dda.dk/.
Fu, Zhichun, H. M. Boot, Peter Christen, and Jun Zhou. 2014. 'Automatic Record Linkage of Individuals and Households in Historical Census Data'. *International Journal of Humanities and Arts Computing* 8 (2): 204–225. https://doi.org/10.3366/ijhac.2014.0130.
[8] Galor, Oded, and Marc Klemp. 2019. 'Human Genealogy Reveals a Selective Advantage to Moderate Fecundity'. *Nature Ecology & Evolution* 3 (5): 853–57. https://doi.org/10.1038/s41559-019-0846-x.
[9] genpop. n.d. 'Genes, Genealogies and the Evolution of Demographic Change and Social Inequality'. GENPOP. Accessed 30 March 2021. http://genpop.org/.
[10] Juel, K., and K. Helweg-Larsen. 1999. 'The Danish Registers of Causes of Death'. Danish Medical Bulletin 46 (4): 354–57.
[11] Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, et al. 2018. 'Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives'. Science 360 (6385): 171–75. https://doi.org/10.1126/science.aam9309.
[12] Mandemakers, Kees, and Jan Kok. 2020. 'Dutch Lives. The Historical Sample of the Netherlands (1987−): Development and Research'. Historical Life Course Studies, June. /article/dutch-lives-historical-sample-netherlands-1987%E2%88%92-development-and-research.
[13] Modin, Bitte, Ilona Koupil, and Denny Vaguero. 2006. 'The Impact of Early Twentieth Century Illegitimacy across Three Generations'. Centre for Health Equity Studies, CHESS, Stockholm University.
[14] Novo Nordisk Fonden. n.d. 'Artificial Intelligence Will Transcribe the Family Relationships of Danes and Strengthen Research'. Novo Nordisk Fonden (blog).

Accessed 12 March 2021. https://novonordiskfonden.dk/en/news/kunstig-intelligens-skal-kortlaegge-danskernes-stam-trae-og-styrke-forskning/.
[15] Ruggles, Steven. 2002. 'Linking Historial Censuses: A New Approach'. History & Computing 14 (1/2): 213–24. 2014. 'Big Microdata for Population Research'. *Demography* 51 (1): 287–97. https://doi.org/10.1007/s13524-013-0240-2.
[16] Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. 'Historical Census Record Linkage'. *Annual Review of Sociology* 44 (1): null. https://doi.org/10.1146/annurev-soc-073117-041447.
[17] Ruggles, Steven, Catherine Fitch, Ron Goeken, J David Hacker, Jonas Helgertz, Evan Roberts, Matt Sobek, Kelly Thompson, John Robert Warren, and Jacob Wellington. n.d. 'IPUMS Multigenerational Longitudinal Panel'.
[18] Ruggles, Steven, Evan Roberts, Sula Sarkar, and Matthew Sobek. 2011. 'The North Atlantic Population Project: Progress and Prospects'. Historical Methods: A Journal of Quantitative and Interdisciplinary History 44 (1): 1–6. https://doi.org/10.1080/01615440.2010.515377.
[19] Song, Xi, and Cameron D. Campbell. 2017. 'Genealogical Microdata and Their Significance for Social Science'. Annual Review of Sociology 43 (1): 75–99. https://doi.org/10.1146/annurev-soc-073014-112157.
[20] Thorvaldsen, Gunnar, Trygve Andersen, and Hilde L. Sommerseth. 2015. 'Record Linkage in the Historical Population Register for Norway'. In Population Reconstruction, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen, 155–72. (online): Springer.
[21] Van Zeeland, Nelleke, and Signe Trolle Gronemann. 2019. 'Participatory Archives'. In Participatory Archives, edited by Edward Benoit III and Alexandra Eveleigh, 103–14. http://ebookcentral.proquest.com/lib/kbdk/detail.action?docID=6031675.

Digital geospatial data, a tool for interpretation of our past.

Gregor Završnik,

Geoarh, Gregor Završnik, s.p., Ljubljana, Slovenija

Abstract. Maps and modern Geospatial records are a tool that has and is used to better understand objects and phenomena around us. With the development of exponential technologies like social networks, Artificial Intelligence geospatial data remains a cornerstone for the development of the digital economy. However, we can only leverage the benefits of this technology if we can assure the discoverability and accessibility of geospatial records in a uniform and accessible way. Today a lot of data is stored in different formats with different levels of documentation and is often only accessible in closed systems.
This paper demonstrates how geospatial data can be used for visualisation and analysis to better understand and use our present and past data using simple and advanced technologies. It discusses what generally brings value to data and what challenges we are facing in the data-driven economy. It then proposes how the Common Information Type Packaging Specifications for geospatial records, developed in the EU eArchiving building block, can support creating an Interoperable and connected information platform that can facilitate innovation and generate new business models. The solution is based on international standards from the Geospatial and Archival domains. The use of eArchiving specification ensures an open and transparent approach that will be sustainable and will ensure legal compliance where that is required.

1 Introduction

Archives and libraries are full of old maps and records describing assets in space. Maps have been used to convey the relationships between objects in space and help us navigate, manage properties, make strategic decisions in war and peace. We determined simple problems like how far a city is from a road to more complex questions like how many bridges are in a specific administrative area to plan their maintenance costs and schedule inspections.

With the computer age, the algorithms for geospatial problem solving got more and more complex. Today, we can use artificial intelligence tools to quickly predict the probability of traffic accidents in a city with great accuracy. Those results are based on interpreting years of geospatial and other data for more than 20 variables. And the longer the time series of data is available, the more precise the predicting models extrapolate.

In the information age, data is the asset, but the real value of an asset is the content itself and the speed by which we can access it. Archives hold considerable amount of heritage data and will also be the keepers of digital records being produced today. Therefore, the future value of the archives will not only be in the data it holds but in the speed of its availability and accessibility. The next multiplier of value is in the connectivity of data from different sources. A telephone network is more valuable if many people use it since you can call almost anyone with the same interface. The value of archival data will be greater if multiple archives can be connected and accessed quickly.

This paper discusses the use cases of geospatial records in interpreting older records and how interoperable geospatial records management can bring value for the preservation and dissemination of modern and heritage geospatial records. Adequately structured, packaged and preserved data is crucial for future applications of the data-driven economy.

Use cases for geospatial records

The best saying that represents the need for spatial data is: "Everything happens somewhere[1]". This statement shows us that the estimate that 80% of all data is related to space or a location has merit.

When people want to locate or manage any objects in space, from a secret location of a buried treasure to large systems like road infrastructure, pipelines or even cities, or countries, maps are used as a platform. Using maps, we can count, measure, compare and analyse different tangible variables, like how many kilometres of roads we need to maintain or where the different soil types can be found.

[1] Durante, Kim, and Darren Hardy. "Discovery, management, and preservation of geospatial data using hydra." *Journal of map & geography libraries* 11, no. 2 (2015): 123-154.

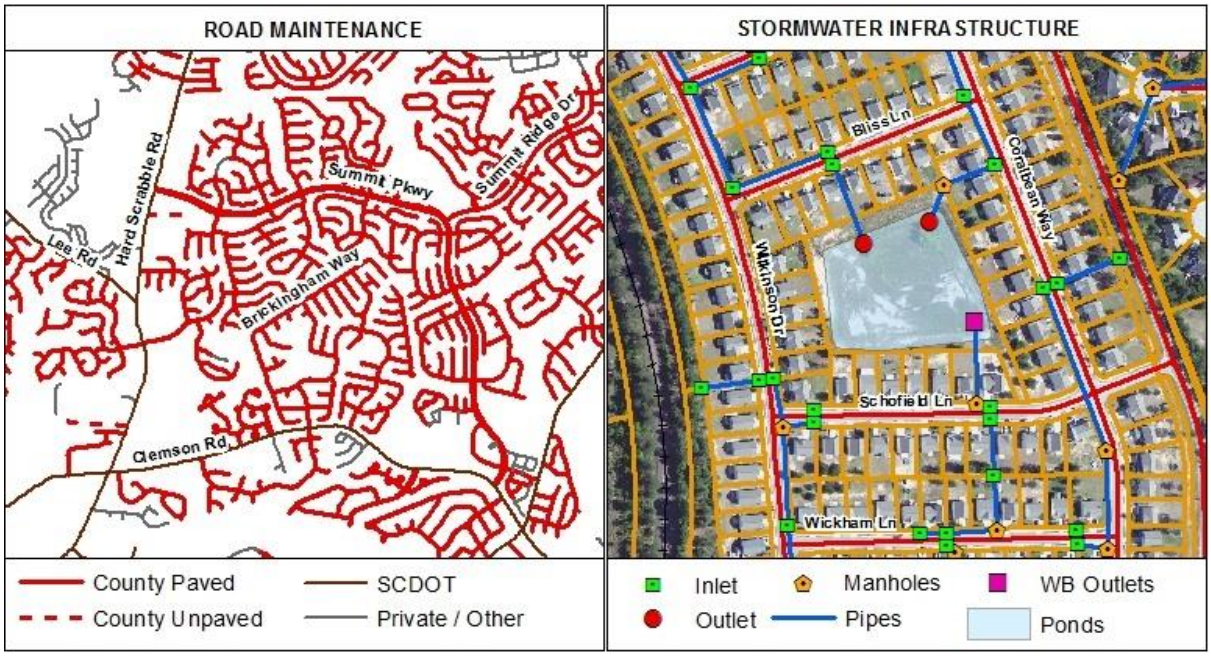


Figure 1 - GIS Usage for Asset maintenance. Source (https://www.richlandcountysc.gov/)

We can also visualise nontangible variables like crime in different statistical regions or compare sales with income ranges across different statistical areas.

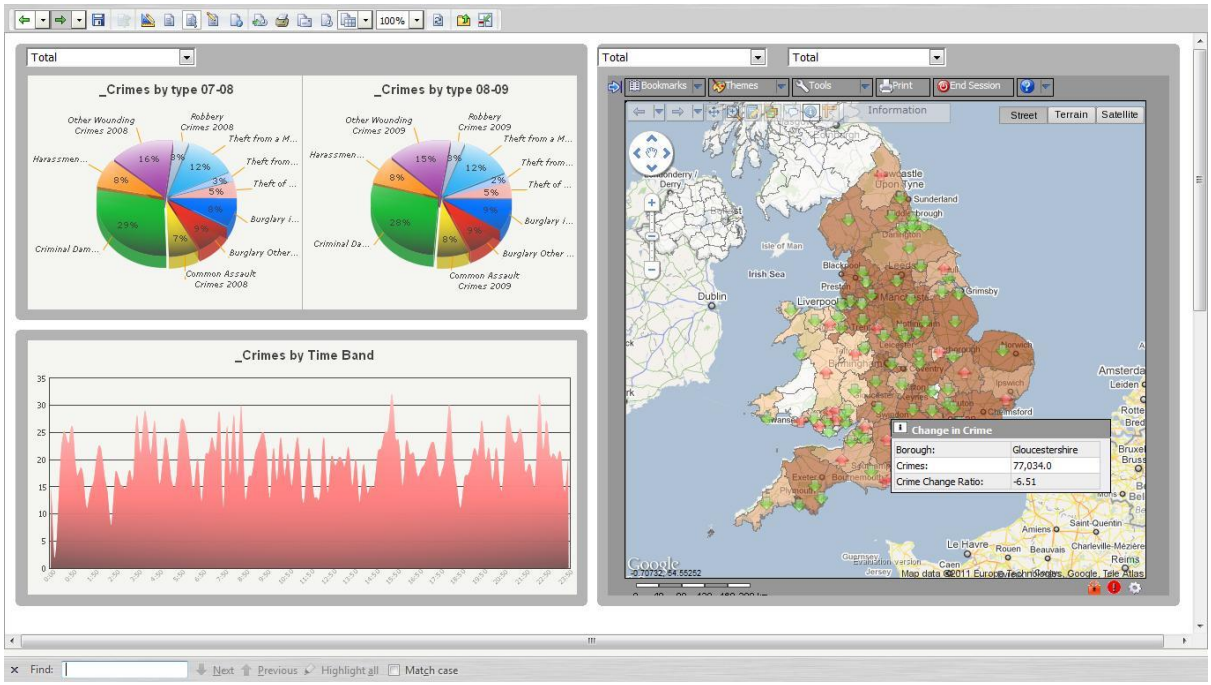


Figure 2 - Crime rate analysis using Business Intelligence tools. Source Microstrategy (www.microstrategy.com)

Geospatial records and GIS

Geospatial Information Systems[2] (GIS) are used to achieve the results mentioned above. With GIS we can capture, manage, transform, analyse and visualise a combination of various mathematically structured geospatial data in different formats[3] combined with other types of digital records.

Some common GIS use cases are:

- Visualisation (2D and 3D)
- Asset management (Management of real estate, cadastral parcels, Linear infrastructures like roads, powerlines, sewage and water)
- Transportation analysis (using for routing, navigation, logistics, etc.)
- Suitability analysis (best location for waste plant locations, a new store, emergency response unit posts, etc.)
- Terrain analysis (analysis of water movements, floods, calculation of volumes of material for excavation)
- Business and government intelligence
- Environmental planning, analysis
- Security planning (military strategy plans, disaster management, etc.)
- Temporal analysis (detection of changes in land cover or voting preferences through time)

Geospatial analysis

Geospatial data represent objects in space, and its mathematically structured nature allows for processing using mathematical algorithms. We can group them in different types of geospatial analysis like visualisation of parameters, overlay analysis, proximity analysis, network analysis, geostatistical analysis, geospatial transformations of data and complex combinations of all types.

[2] Wikipedia. "Geographic Information System", Accessed August 30, 2021. https://en.wikipedia.org/wiki/Geographic_information_system

[3] Wikipedia. "GIS File Formats", Accessed August 30, 2021. https://en.wikipedia.org/wiki/GIS_file_formats

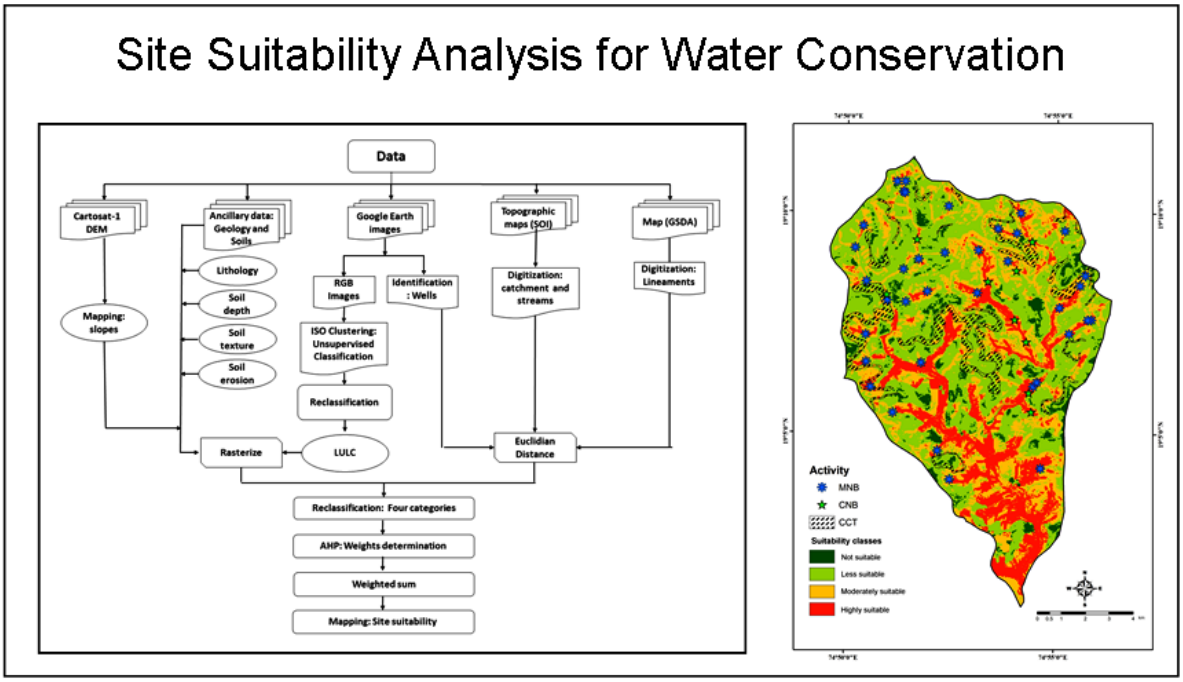


Figure 3 - Suitability analysis example - Algorithm and the resulting map

Geospatial data and Heritage records

Geospatial Analysis was conducted long before the arrival of modern computers. One of the first use cases was the epidemiological study of cholera by John Snow in 1854[4]. We can see plans of battles on old maps, and we know that cadastral maps were used for collecting taxes.

Maps were also used to depict results of analysis of social phenomena like Frequency of marriages and much more.

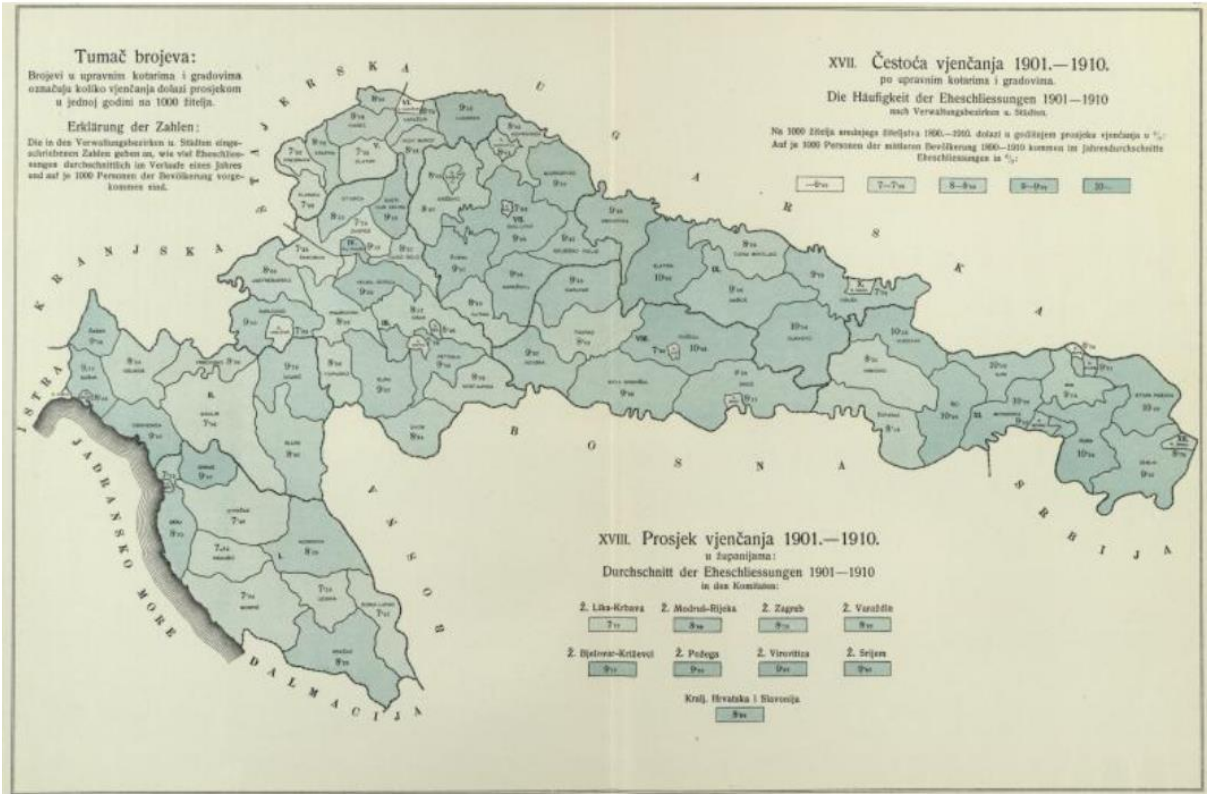


Figure 4 - Analogue GIS Usage - Map of Frequency of marriages per 1000 persons between 1901-1910 in Croatian administrative units of Croatia and Slavonija

Adding context to Heritage records with geospatial data
With today's technology, we can use GIS functionality of analysis and 2D and 3D visualisation to add additional context to heritage records and bring them to a wide audience faster than ever before.

The challenge is the process of digitisation of analogue records that costs time and resources. However, there are possible shortcuts with the modern tools of the digital economy like Machine Learning, Crowdsourcing and Gamification.

[4] Brody, Howard, Michael Russell Rip, Peter Vinten-Johansen, Nigel Paneth, and Stephen Rachman. "Map-making and myth-making in Broad Street: the London cholera epidemic, 1854." *The Lancet* 356, no. 9223 (2000): 64-68.

An example use case could be a possible extension of an ongoing project led by dr. Hrvoje Stančić, at Department of Information and Communication Sciences, University of Zagreb, Croatia. In this project, machine learning is used to automate digitising Food Rationing cards from World War 2.

In the project, street names were one of the fields recognised from the Food Rationing cards. We discussed extending the project by georeferencing some data from the cards, like socioeconomic data (income, religion), which would give us additional context to the researched period.

Here is where we meet the challenge of identifying the old addresses with the new ones. And this is where we see that the analogue format and the lack of connectivity between the analogue records in the way we are used in the digital realm are preventing us from using the old data in the ways we use current data.

The georeferencing could be done by matching old addresses with the new ones, but if data didn't exist, a gamified crowdsourcing campaign could be launched with schools to match old ones with new ones. And this brings us to the exponential technologies of the data-driven economy.

What brings value to data in a Data-driven economy?

New technologies are bringing information to users faster than ever before. That is why the speed of adoption is faster today than it ever was, and new innovations can be introduced faster.

A data-driven economy is becoming the main focus of all major economic powers. That is one of the reasons that the EU has launched the Digital Europe Programme[5], where the aim is to promote the development of what could be called exponential technologies (AI, Supercomputing, wider use of digital technologies across economy and society).

By providing context and space as a common denominator, geospatial data was also mentioned as the cornerstone of the digital economy[6] and can therefore be considered an essential contributor to bringing value to all other data.

The familiar quote: "*The knowledge is power*" is known from almost half a century ago, when Sir Francis Bacon (allegedly) mentioned it in his published work Meditationes Sacrae (1597). And it comes from a time when the knowledge was scarce and unattainable.

However, today the situation has changed since we have so much information available that we can't process it in a meaningful way. The quote still has some merit, and it points us in a useful direction, but due to the changed availability of information lacks specificity. Now we should say, "*Applied knowledge is power*". And the trick is in the word applied.

So, let's ask ourselves how the data from the records we can offer could be applied best. We come to a couple of new variables that influence value: discoverability, availability and speed of access.

SCIENCE author Steven Johnson once said: "innovation doesn't come just from giving people incentives; it comes from creating environments where their ideas can connect."[7] And this is a direction that will bring the most value.

The question then is: How can we make the data discoverable, available and quickly accessible?

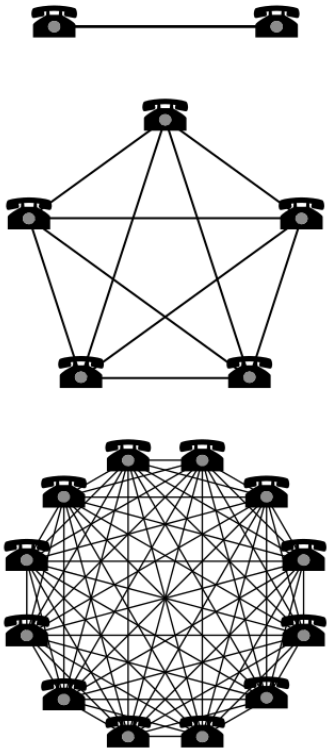
[5] European Commission. “The Digital Europe Programme”, Accessed August 30, 2021. https://digital-strategy.ec.europa.eu/en/activities/digital-programme

[6] European Commission. “Building Data Economy Brochure”, Accessed August 30, 2021. https://digital-strategy.ec.europa.eu/en/library/building-data-economy-brochure

[7] Steven Johnson, Where Good Ideas Come From: A Natural History of Innovation (N.Y.: Riverhead, 2010).

Connectivity - The power is in numbers

In the field of telecommunications, they use the Metcalfs Law[8] to determine the value of a communications network. And it states that the value of a (tele)communications network is proportional to the square of the number of connected users of the system (n^2). The term end-users was referring to compatible communicating devices as opposed to people.



And this makes sense. The more people that are connected to the network, the more useful it is. We can use only one network to contact almost anyone, and it saves us time from changing between means of communication.

If we use this analogy to use data that Archives can offer, we can quickly see that the value of a network of Archives rises if the end-users can have fast and uniform digital access to any archive. And since the end-users of the digital economy are mostly going to be automated systems, the value of a data repository will be uniform compatibility.

And this is what brings us to the need for ***interoperability***[9] among all members of the communication network if we want to ensure a valuable platform serving the Data Economy.

Figure 5 - Metcalf's law - Value of a communication network is in number of connections, which grow exponentially

[8] Wikipedia. "Metcalfe's Law", Accessed August 30, 2021. https://en.wikipedia.org/wiki/Metcalfe%27s_law
[9] Wikipedia. "Interoperability", Accessed August 30, 2021. https://en.wikipedia.org/wiki/Interoperability

Interoperability of Geospatial Preservation records

When organisations started using GIS, it was common that every organisation developed their own system with their own formats. Eventually, the geospatial community discovered that their data is only as valuable as the possibility of interchanging it with others.

This awareness led to the development of geospatial standards and standardised formats, interfaces between systems and standardised metadata descriptions to facilitate the discoverability of the data. OGC[10] is a global organisation created by more than 500 businesses, government agencies, research organisations, and universities driven to make geospatial (location) information and services FAIR - Findable, Accessible, Interoperable, and Reusable.

On the European level, interoperability is achieved for the foundational geospatial datasets by the INSPIRE directive[11]. As described on their web page: "*INSPIRE directive aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. This European Spatial Data Infrastructure will enable the sharing of environmental spatial information among public sector organisations, facilitate public access to spatial information across Europe and assist in policy-making across boundaries".* But to ensure this level of interoperability between nation-states, a standardised data model is prescribed for a limited number of datasets.

However, most of the initiatives only address the current geospatial records. They do not deal with the old geospatial records or consider the long-term preservation aspects of new records.

That is why the E-ARK Project[12] was initiated by the EU archives and other interested parties and started in 2014. The project aimed to develop internationally accessible archives through the provision of technical specifications and tools, the development of an integrated archiving infrastructure, the demonstration of improved availability, access and use, and the rigorous analysis of aggregated sets of archival data.

The success of the E-ARK project resulted in the adoption of the developed technical specifications and tools by the CEF eArchiving[13] building block. Under eArchiving, the technical specifications were further developed to facilitate the interoperability of digital records preservation across the EU digital market.

[10] Open GIS Consortium. “About OGC”, Accessed August 30, 2021. https://www.ogc.org/about

[11] European Commission. “About INSPIRE”, Accessed August 30, 2021. https://inspire.ec.europa.eu/about-inspire/563

[12] E-ARK Consortium. “E-ARK Project”, Accessed August 30, 2021. https://eark-project.com/

[13] European Commision. “eArchiving”, Accessed August 30, 2021. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving

eArchiving – Interoperability, openness and sustainability

So, what is eArchiving? It is a Building Block within the Common European Facility[14] (CEF) intended for anyone whose information should be kept accessible and reusable for years to come, regardless of the system used to store it. To achieve it, eArchiving provides core specifications, software, training, and knowledge to help people preserve and reuse information over the long term.

Interoperability is built into eArchiving since the technical specifications were built based on international standards. And having a common set of open specifications for packaging and archiving digital information promotes a high level of transparency and confidence among all participants in the information lifecycle. With eArchiving, digital archival systems can be successfully deployed, implementing reusable modular components compliant with various national legal backgrounds.

The technical specifications are now maintained by the Digital Information LifeCycle Interoperability Standards Board[15] (DILCIS Board). DILCIS Board is an international group of experts committed to maintaining and sustaining a set of interoperability specifications that allow for the transfer, long-term preservation, and reuse of digital information regardless of the origin or type.

Common Specification for Information Packages (CSIP) - the cornerstone of eArchiving

The Common Specification for Information Packages (CSIP) aims to serve three primary purposes:

- Establish a common understanding of the requirements which need to be met to achieve Interoperability of Information Packages;
- Establish a common base for the development of more specific Information Package definitions and tools within the digital preservation community;
- Propose the details of an XML-based implementation of the requirements using standards widely used in international digital preservation to the largest possible extent.

Ultimately the goal of the CSIP is to reach a level of Interoperability between all Information Packages so that institutions can take up tools implementing the CSIP without needing further modifications or adaptations.

[14] European Commission. “CEF Digital Home”, Accessed August 30, 2021. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home

[15] DILCIS Board. “The Digital Information LifeCycle Interoperability Standards Board”, Accessed August 30, 2021. https://dilcis.eu/about

The data model structure is based on a layered approach for information package definitions. The Common Specification for Information Packages (CSIP) forms the outermost layer.

The general Information Package specifications (SIP, AIP and DIP[16]) add a submission, archiving and dissemination information to the CSIP specification.

The third layer of the model represents specific Content Information Type Specifications (CITS). The CITS for Geospatial is specifying what and how geospatial records should be preserved.

Additional layers for business-specific specifications and local variant implementations of any specification can be added to suit the needs of the organisation.

CITS for Geospatial[17] – Guideline for Interoperability of preservation and reuse of Geospatial records

The CITS Geospatial specification describes how geospatial data files, metadata files, schema files for validation, documentation, and other files should be placed and structured into the CSIP based structure when producing a CITS Geospatial SIP for transfer to long-term preservation and from a repository to dissemination.

The specification is general enough to support multiple types of geospatial records (not only vector and raster-based records). Therefore, the specification does not define mandatory long-term preservation formats. Instead, it provides a possibility of extensions, the so-called Long-term preservation format Profiles, that need to comply with general requirements. Examples of such Profiles for vector data (GML[18]) and raster data (GeoTIFF[19]) are provided in the two guidelines accompanying the CITS document. Profiles for other geospatial record formats (like proprietary data, earth observations, point clouds, oblique images, web services, etc.) are not proposed at this stage. They will be added later in cooperation with the geospatial and preservation community.

CITS Geospatial package structure

The Content Information Type Specification for Geospatial data aims to define the necessary elements required to preserve the accessibility and authenticity of geospatial records over time and across changing technical environments. To achieve it, this

[16] Society of American Archivists. »Open Archival System (OAIS), Accessed August 30, 2021. https://www2.archivists.org/groups/standards-committee/open-archival-information-system-oais

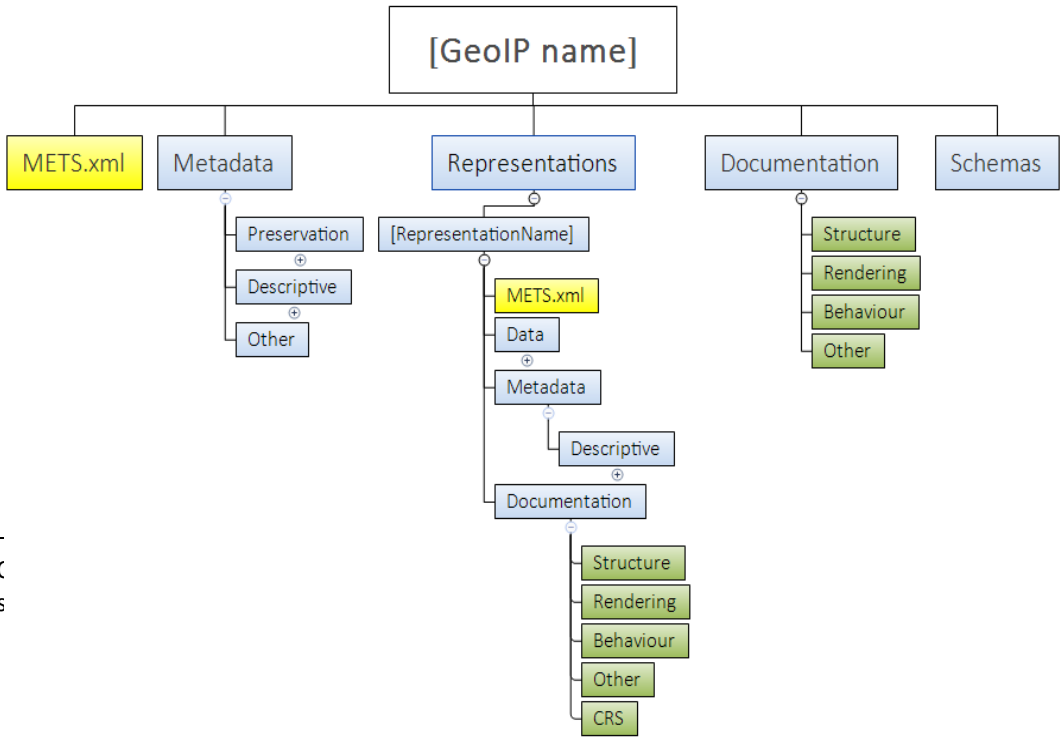
[17] DILCIS Board. "CITS Geospatial / Specification", Accessed August 30, 2021. https://github.com/DILCISBoard/CITS-Geospatial/tree/master/specification

[18] Wikipedia. "Geography Markup Language", Accessed August 30, 2021. https://en.wikipedia.org/wiki/Geography_Markup_Language

[19] Wikipedia. "GeoTIFF", Accessed August 30, 2021. https://en.wikipedia.org/wiki/GeoTIFF

specification defines the categories of significant properties[20] for geospatial records to allow the digital geospatial information products to remain accessible and meaningful. For every geospatial record or a set of records, we need to preserve information that suits the following categories:

- **Content** – Information contained within the Information Object. For example, location information (coordinates, orientation, pixel size), geometry, related feature attributes, etc.
- **Context** – Any information that describes the environment in which the content was created or that affects its intended meaning. Examples are: Creator name, date of creation, spatial accuracy, source data, sensor information, etc.
- **Structure** – Information that describes the extrinsic or intrinsic relationship between two or more types of content, as required to reconstruct the performance. For example, a Raster object and its connection to the world file, or a vector dataset combined with a table, a GIS project, defining the structure of geodata layers used to create a map, configuration of web services, defining a mash-up WMS, etc. This information should preferably be provided using standardised machine-readable files or at least in a written documentation.
- **Rendering** – Any information that contributes to the recreation of the performance of the Information Object. Example: Colour map of pixel values of a raster; Styled Description layer for web services, documentation describing a cartographic map project, Report designs, etc.
- **Behaviour** – Properties that indicate the method in which content interacts with other stimuli. Example – rendering algorithms, analysis functionalities, standard transformation processes, documentation of original system functionality, user manuals, training materials, system usage videos, etc.



[20] InSPEC
https://s

CITS Geospatial validation requirements

A general Information Package structure is defined with CSIP requirements. CITS Geospatial builds on CSIP and adds additional requirements specific to Geospatial records.

The CITS Geospatial specification is structured in sets of criteria that a package must fulfil to be a valid CITS Geospatial package. The criteria contain an ID designation, Description with location and cardinality level for the requirement (MUST, SHOULD, MAY).

ID	Name, Location & Description	Card & Level
GEO_11	Minimum one file in a geospatial format If the value in mets/@csip: CONTENTINFORMATIONTYPE is "GeoData", then there **SHOULD** exist at least one file in a geospatial format in representations/[RepresentationName]/data	0..n SHOULD

Figure 6 - CITS Geospatial extension folders for Information Packages

GEO_12	Subfolders in data representations/[RepresentationName]/data If there are more geospatial records in a representation, each geospatial file **MAY** be placed or grouped in subfolders in representations/[RepresentationName]/data	0..n MAY

Table 1 - Example of validation requirements for CITS Geospatial

The validation criteria are divided into five major groups:

Folder structure requirements
This set of criteria defines additional folder structure requirements and options for storing certain types of elements that need to be in a Geospatial SIP. The designation is intended for easier machine-readable access.

METS Requirements
The CITS Geospatial extends the requirements for METS[21] metadata, specific to geospatial records. This part intends to describe if the Information Package or a part of it contains Geospatial records.

[21] Library of Congress. "METS Metadata Encoding & Transmission Standard" Accessed August 30,2021. https://www.loc.gov/standards/mets/

Data Requirements
This chapter states the requirements for the content data object or objects that form the geospatial record contained in the Information package. It also provides specific requirements for vector and raster data types and defines what a Long term preservation Format Profile should contain.

Documentation requirements
Geospatial records are very complex and often require additional technical and descriptive documentation for us to be able to reuse them accordingly. This chapter defines what is needed to define its structure, rendering, behaviour and other documentation. Some of these elements need to be in standardised machine-readable formats and placed in designated locations to enable automated reuse.

(Geospatial) Metadata Requirements
This chapter is referring to requirements for geospatial metadata content. This content is crucial to facilitate automated discovery using existing Geospatial Metadata Catalogues. It supports standardised metadata like INSPIRE, ISO 19115 and others.

Guidelines for Geospatial Records and GIS systems
The CITS Geospatial comes with the two accompanying guidelines that aim to explain how to use it for the preservation of Geospatial records and GIS systems:

- The first accompanying guideline document (Guideline for the specification for the E-ARK Content Information Type Specification for Geospatial data (CITS Geospatial)) provides a basic introduction to the field of geospatial data and the concepts used in this specification. In the guideline, there is also a table translating metadata elements from INSPIRE directive metadata into Archival metadata standards (ISADG and EAD3).

- The second guideline document (Guideline for using the specification for the E-ARK Content Information Type Specification for Geospatial data (CITS Geospatial) with GIS) provides the information on how to extend the first accompanying guideline document with content describing preservation of selected elements from Geographical Information Systems (GIS). The guideline aims to extend the scope of preservation beyond the geospatial data records themselves and focus more on GIS elements defining the geospatial information products.

Conclusion
As established, geospatial data is an important ingredient for understanding the objects and social phenomena around us. The visualisation and analytic capacity of Geographic Information Systems provide tools for a better understanding of our past, which helps us make better and more informed decisions for the future.

The value of all data, geospatial or not, is in its usability, which can be further defined by terms like discoverability, accessibility, speed of access and connectivity with other databases.

We can support future innovations in business models by providing a connected platform of interoperable and accessible data sources. That is how the future users, human or machine, can access available geospatial and other data seamlessly and interoperably and develop new disruptive products and services that were previously unavailable.

By adopting readily available eArchiving standards for uniformed packaging, documenting and preserving Geospatial and other data types, data repositories such as archives, data producers, and solution providers will also increase their value as members of this connected platform. Can ensure uniform access to this data with future users across Europe.

Conclusions

Zoltán Szatucsek

Senior Archivist in the National Archives of Hungary, Director of the Innovation and IT Department

We all have heard about the miracles that we can read books without open them, we can build a family tree automatically and so on, and during these two exciting days we wanted to find out the truth about this.

It became clear that these miracles are in different stages of completion. We heard about successful projects, ongoing developments and ambitious visions shared by some speakers.

Kerstin Arnold reported on how the APE portal is trying to automatically create search aids in a multilingual environment.

Artem Reshetnikov's spectacular presentation showed **how important** cultural and historical context is in using computer vision.

We came closer to understanding **what is easy and what is difficult** for scanning with non-invasive technologies. Mass digitisation wont be solved by X-Ray CT, but it can be important in the preservation of individual documents.

A promising project by colleagues at Humboldt University **foresees** that through the description of coats of arms --- we can gain further insights into the historical connections between different areas of Europe. Gregor showed how geodata can bring new context to old records. Lipót presevation was very convincing about technology sometimes help reduce time and costs for projects besides being more accurate than humans.

The Danish Link lives project is creating the lifecourses of individuals **from databases that were created decades earlier**, re-used now in new and innovative ways. But Joan Andreu and Enrique Vidal showed that recognising the handwriting of images that still have not yet been indexed is not an impossible task anymore.

In fact, the EDT project has already done this on 75,000 pages of names. Combining the two technologies opens up incredible perspectives.

The most important messages of these two days were highlighted by the panelists. Archives need to redefine, reconceptulize themselves. We don't manage records, we manage information. New, disruptive technologies - handwriting recognition, machine vision, machine learning, geodata management - are helping us to extract this information, process it and make it available to our users.

The school year is starting in each countries these days. **Sooo,** to be stylish we have a lot of homework.

- First of all, we must be patient. Developments are at different stages of completion and there no miracles exist.
- We have to be curious. Archivists are still key persons in the future, but as Kerstin said, we have to ask questions and remain curious and open minded.
- Large institutions have a key role to play in this, just as they cannot leave smaller institutions alone.
- And finally we need to support researchers and developers with good quality and sufficient quantity of data. It is a mandatory precondition for research.

The intention of these two days was to understand how the magic of new technology works. There is no ready-made, out-of-the-box product, but the presentations were invariably inspiring and opened up fundamentally new perspectives and encouraged us to refocus our activities.

I have nothing more to do just to wish you all a successful start to the school year **or much more years**.

Thanks for the excellent organization. Severiano, Cristina, Miguel and the partners of the EDT and Maria for hosting us now in Alicante. I hope you enjoyed. Have a safe trip to home.