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

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

1 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.

2 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.

1 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.

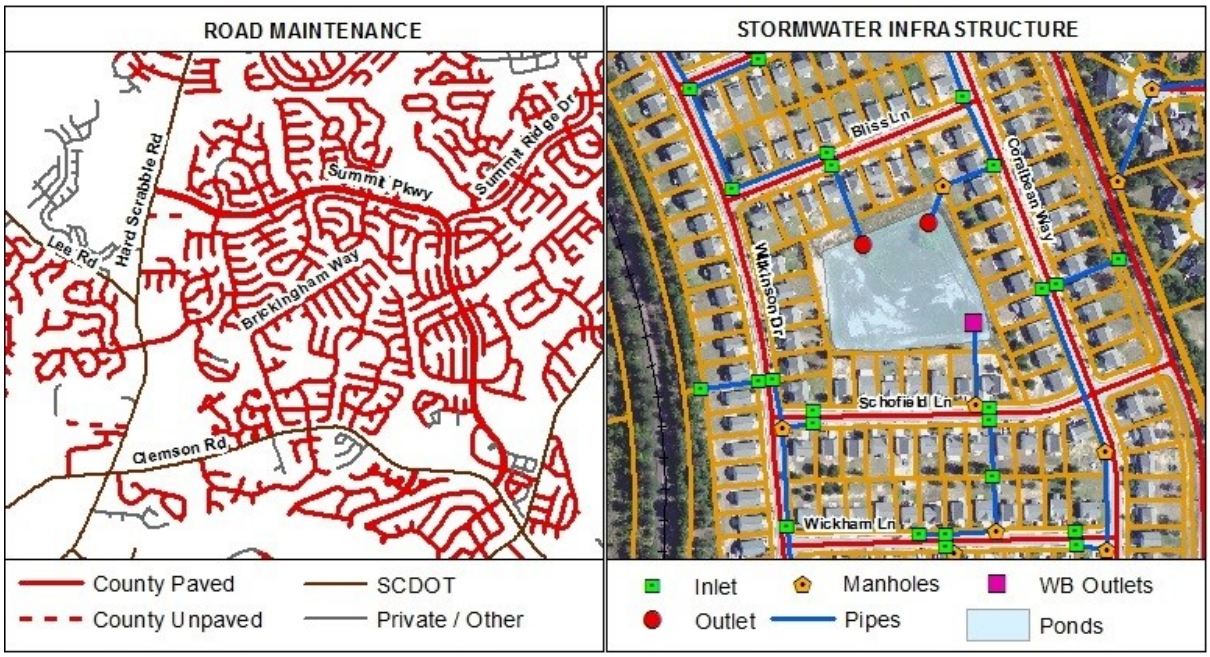

Figure 1 - GIS Usage for Asset maintenance. Source (https://www.richlandcountysc.gov/)

[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.

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

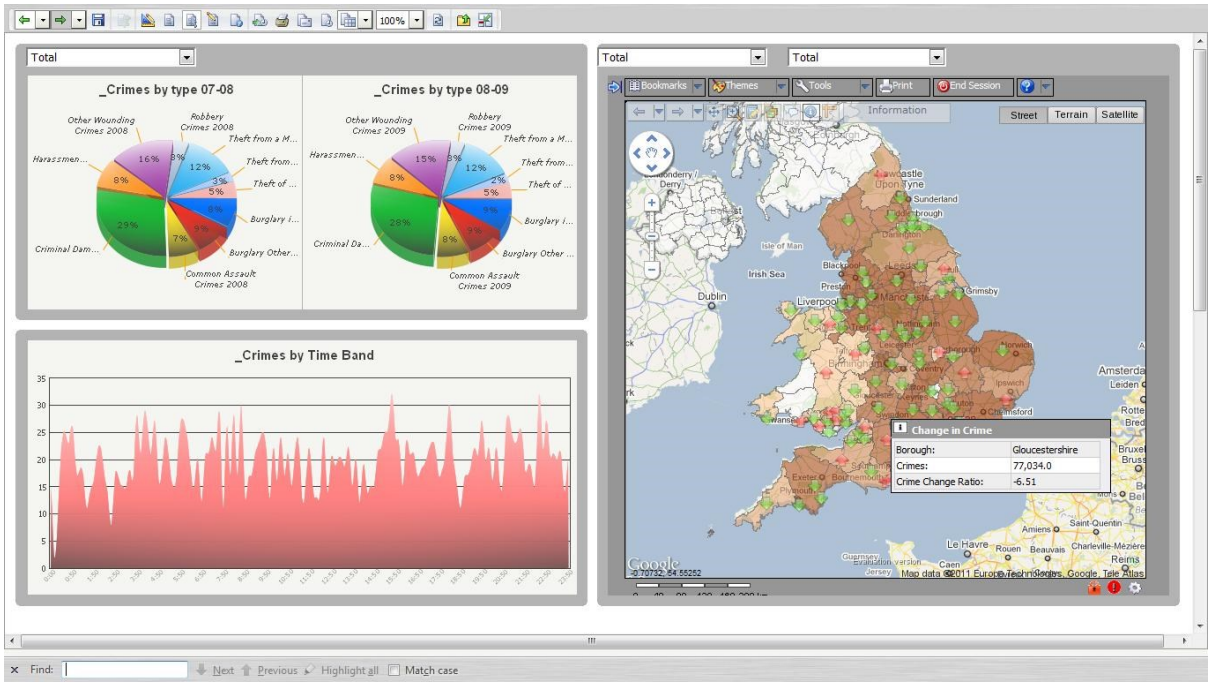

Figure 2 - Crime rate analysis using Business Intelligence tools. Source Microstrategy (www.microstrategy.com)

1.1 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

[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

- Security planning (military strategy plans, disaster management, etc.)
- Temporal analysis (detection of changes in land cover or voting preferences through time)

1.2 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.

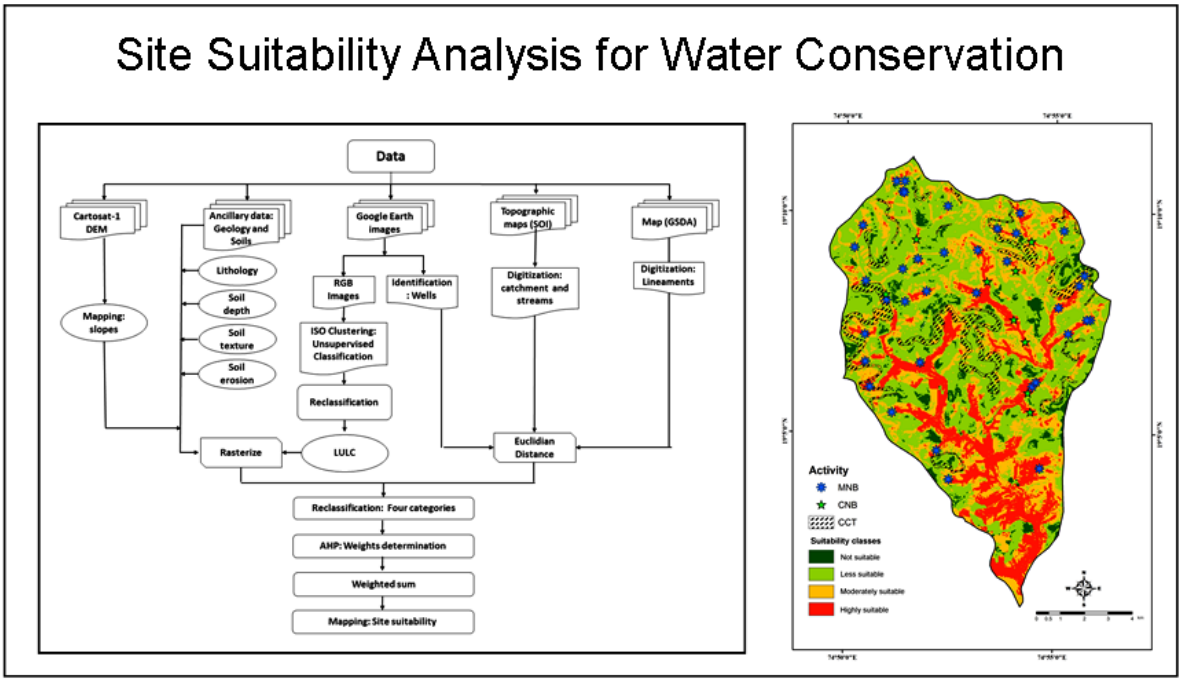

Figure 3 - Suitability analysis example - Algorithm and the resulting map

1.3 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.

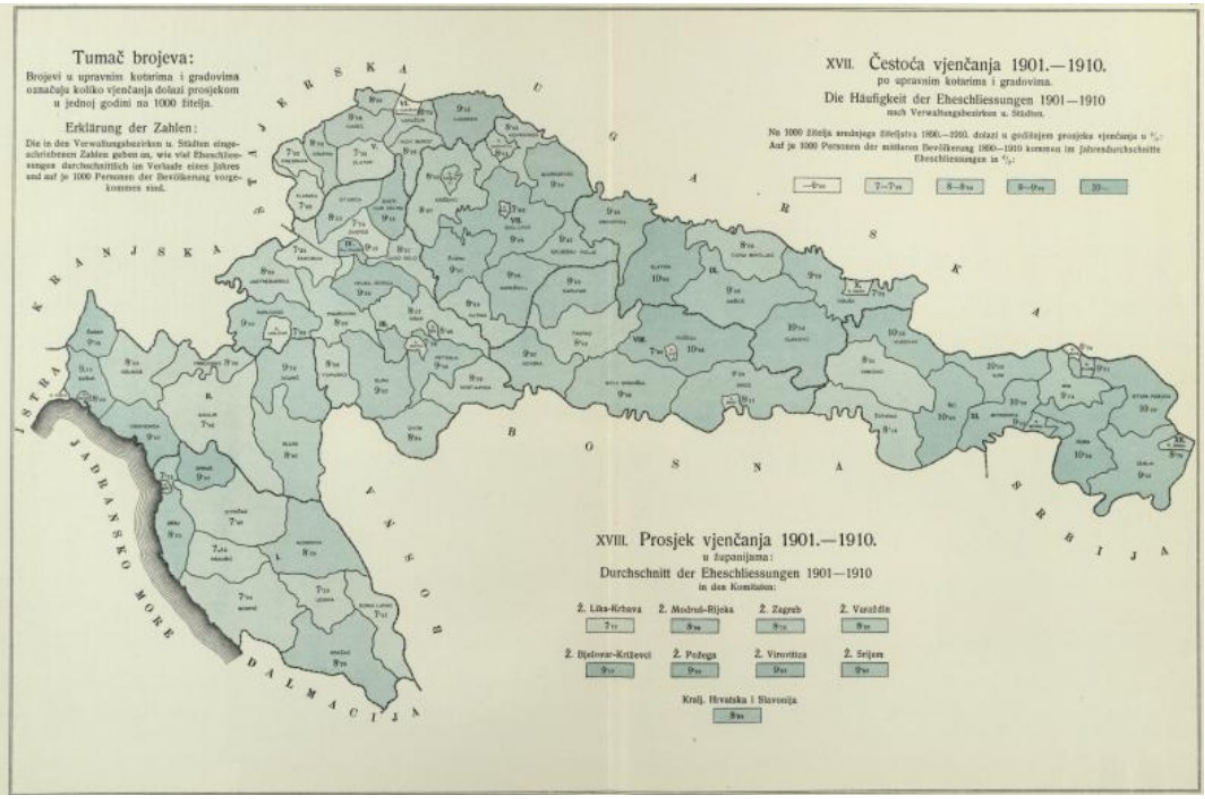

Figure 4 - Analogue GIS Usage - Map of Frequency of marriages per 1000 persons between 1901-1910 in Croatian administrative units of Croatia and Slavonija

1.3.1 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.

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,

[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.

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.

2 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).

2.1 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.

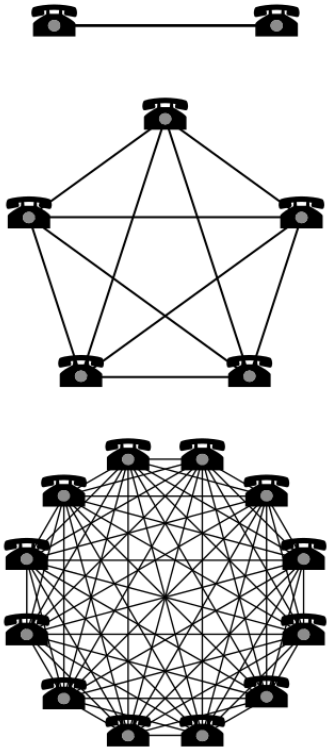

Figure 5 - Metcalf's law - Value of a communication network is in number of connections, which grow exponentially

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.

[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

3 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

3.1 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.

3.1.1 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.

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.

---

[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 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.

3.2 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.

3.2.1 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 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:

---

[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

[20] InSPECT Project. Significant properties and digital preservation", Accessed August 30, 2021. https://significantproperties.kdl.kcl.ac.uk/

- **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.

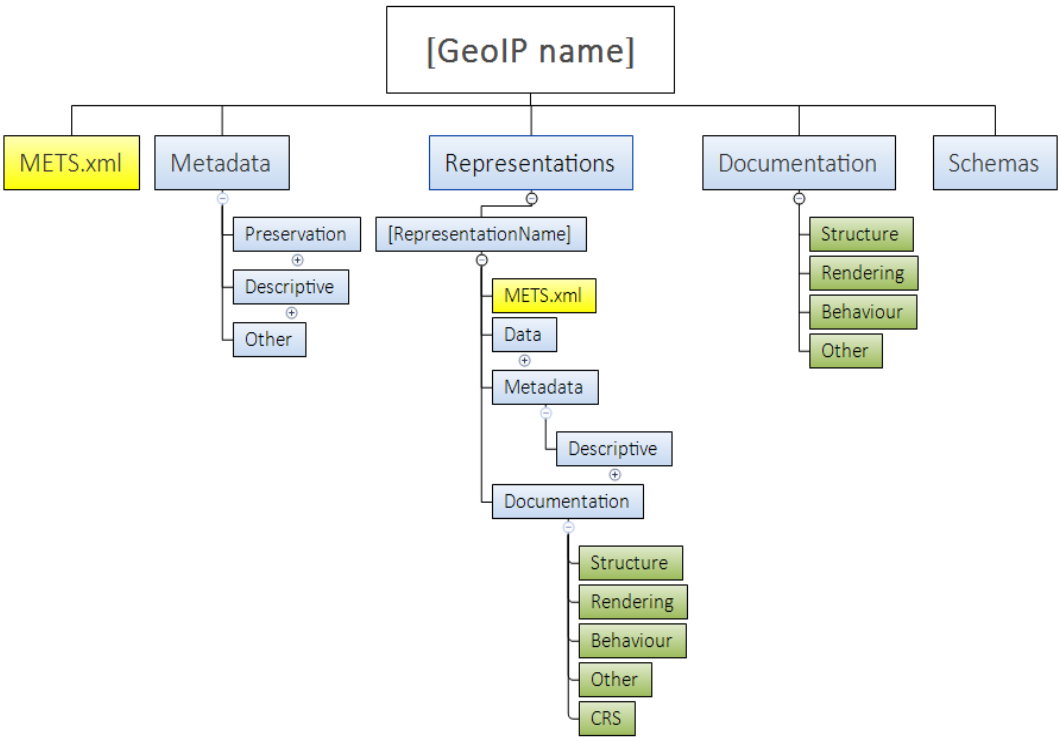

Figure 6 - CITS Geospatial extension folders for Information Packages

3.2.2 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
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:

3.2.2.1 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.

3.2.2.2 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.

3.2.2.3 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.

3.2.2.4 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

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

elements need to be in standardised machine-readable formats and placed in designated locations to enable automated reuse.

3.2.2.5 (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.

3.2.3 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.

4 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.