GI+100: Long term preservation of digital Geographic Information — 16 fundamental principles agreed by National Mapping Agencies and State Archives

by Carsten Rönsdorf, Paul Mason and Jonathan Holmes, Ordnance Survey; Urs Gerber and André Streilein, swisstopo; Marguérite Bos, Schweizerisches Bundesarchiv; Arif Shaon, Rutherford Appleton Laboratory; Kai Naumann, Landesarchiv Baden-Württemberg; Michael Kirstein, Generaldirektion der Staatlichen Archive Bayerns; Göran Samuelsson, Mid Sweden University; Marja Rantala, Maanmittauslaitos; Sidsel Kvarteig, Statens kartverk; Lynne Ralsberg and Jenny Svennewall, Lantmäteriet and Wolfgang Stößel, Landesamt für Vermessung und Geoinformation Bayern.

Executive Summary

This paper states 16 principles for the long term retention and preservation of digital geographic information. The paper is mainly aimed at public sector geographic information providers in Europe (particularly those involved in mapping and cadastre) with the intention of highlighting the significance of fundamental concepts for digital geographic data archiving. Geographic information providers are mainly mapping agencies, but also archives preserving geographic data among a wider range of digital information. A supplementary objective is that the paper may provide useful information for providers of all types of geographic information right around the world.

There are many reasons why people wish to retain access to information, though the main drivers for archiving digital geographic information are meeting legislative requirements, the short and long term exploitation (re-use not only access) of archived data for analyzing social, environmental (e.g. global climate changes) and economic changes over time as well as efficiency savings in managing superseded datasets. This paper sets out the path and describes what needs to be done now to future-proof the investment government agencies around the world have made in creating digital Geographic Data. It was approved by the EuroSDR Board of Delegates on 30th May 2013 and also approved by EuroGeographics at their General Assembly on 1st October 2013 and the European Board of National Archives at their General Assembly on 15th November 2013.

The principles are:

1. Archiving of digital Geographic Information begins at the point of data creation, rather than at the point of withdrawal from active systems
2. All organisations must have a maintained Archiving Policy
3. Be selective and decide what to archive and what to dispose of
4. Consider preservation timeframes of 1, 10, 100 years
5. Migration or emulation is inevitable in the medium and long term. Be prepared and choose which properties to preserve in advance
6. The output of the archival planning process should also be preserved over the long-term to accommodate future preservation requirements
7. Archiving is not back-up. You must also back-up your archive
8. Geographical data should be preserved in a way that non geo-specialists can handle
9. Information objects should be self-contained and independently understandable
10. Keep the gold copy version of the 100 year data archive in open, file based repositories, not in databases, nor other complex environments
11. Consider keeping a graphical representation alongside the logical representation of the data
12. Restrict the number of formats and encodings to a widely agreed set of open, simple and well-documented file formats
13. Prefer simple data models and schemas over complex ones
14. Keep the access mechanism for archived data simple. Focus on basic current user requirements – an archival viewing system does not need to be a fully functioning GIS
15. Ensure effective management and quality assurance of the metadata associated with your data
16. Make some assumptions about future use, but don't be too restrictive

Introduction

Looking back 100 years we find a lot of geographic information created at that time that is still very accessible and usable in the form of paper maps. Paper has proved itself a good medium for long term preservation if it is looked after with appropriate care. It is easy to handle and information in graphical form is immediately accessible. In today's digital world however, geographic information is produced predominantly in digital formats that have a very strong reliance on technology to both store and access data. Modern data structures, geometries and data relationships are undoubtedly a lot more powerful and accurate than what, historically, could be drawn on a paper map alone. However, the risk of losing data can be much higher, especially if no precautions are taken. If we look 100 years into the future it is quite difficult to see how the preservation of digital data will be assured and how the information we generate today will be accessed.

Current geographic information that depicts the world we live in today provides tremendous value for government, businesses and research purposes as well as the general public. At the same time historic information is also frequently accessed and adds to this cumulative value. It is our understanding that safeguarding today's fundamental geographic data for future generations in order to understand history as well as historic trends should be a core objective for National Mapping Agencies (NMAs), Archives and other data producers and providers.

While nobody will be able to accurately predict what technology will look like in 100 years time, there is a lot we can do now to prepare so that we can mitigate the risk that datasets will lose their value over time. The likelihood that any technical or organisational assumptions or predictions we make over a 100 year time-frame do not come true is very high. However, the spirit of this paper is that it is sensible to make certain assumptions today and concisely document these to enable future generations to understand today's motivations and decisions. As a result we call on digital data producers and providers to start future-proofing their data now in order to create sustainable long-term resources.

The document has been created by the EuroSDR data archiving working group of 11 National Mapping Agencies, National Archives and Research Institutions across Europe in order to seek endorsement of the set of principles by EuroSDR and EuroGeographics as well as the Archiving community i.e. the EBNA (European Board of National Archivists) / EAG (European Archives Group) and the EURBICA (European branch of the International Council of Archives). It is our goal to establish common principles and a common approach on archiving geographic information. Future guidance by the EuroSDR data archiving working group is expected to be published at http://tinyurl.com/GI100.

Glossary

There is often confusion between terms used in the geographic and archiving communities. Terms used within this document have been used with the following meanings.

Geographic Information

Geographic information is:

- information about places on the Earth's surface
- knowledge about where something is
- knowledge about what is at a given location

(Source Goodchild 1997)

Archiving

The act of transforming administrative records into archival records thus creating the conditions for historical and social research. Records can be all kind of authentic information created by persons or organisations. Archiving encompasses the input, appraisal, description, preservation and provision of documentation and contributes to a secure basis for law, as well as to continuity and efficiency in

administration. Historical archives are responsible for keeping selected historical records for unlimited time. Archiving usually occurs when the primary use of records has fallen below a certain limit.

Format migration

Format migration is the process in which data is migrated from one format to another, to avoid the former becoming obsolete.

Media migration

Media migration is the process in which data is migrated from one archiving media to another e.g. CD -> DVD or CD to tape.

Digital Geographic Data Archiving Principles

The EuroSDR data archiving working group has identified and agreed upon a set of common and practical fundamental concepts and principles related to archiving Geographic Information (GI). These are listed in this chapter and are intended to establish the foundation for public sector providers of GI and Archives around Europe to preserve important GI for future use. It should be noted that legal mandates may mean that these principles cannot be followed and that organisations are legally obliged to divert from the principles and concepts documented below.

Outside the GI domain, more generic and comprehensive conceptualisations of an archive lifecycle already exist. All of these concepts have been heavily influenced by the Open Archival Information System (OAIS, ISO 14721[1]). For example, the DCC Curation Lifecycle Model has been designed to facilitate a lifecycle approach to the management of digital materials in an archive and to enable their successful curation and preservation, including initial selection for reuse and long-term preservation (Higgins, 2008).

The order of the principles follows the lifecycle of data from design and creation to maintenance and preservation (ingestion, annotation, migration), as well as accessing archived data. Suggested action points and important questions where further investigation may be required are indicated by this symbol: ►.

Principle 1: Archiving of digital Geographic Information begins at the point of data creation, rather than at the point of withdrawal from active systems.

Today archiving is often seen as an afterthought, though the long term value of a dataset can often be appraised at the outset. If this appraisal occurs, archival requirements are clear from the start and can be acted upon.► Consult early with those responsible for and proficient in long-term preservation of digital data, i.e. Archives, Libraries or Data Centres, depending on national legislation.

► Define whether long term preservation is desired or necessary.

► Determine and document the retention period. This can be changed at a later date if requirements change but will clarify archival needs from the outset. Such an appraisal should be undertaken for all existing datasets.

Principle 2: All organisations must have a maintained Archiving Policy

► The backbone for any archiving business case is the establishment and agreement of a common preservation planning process and a set of common preservation objectives between data producers and archives.

► When setting up an archive look across borders and beyond its currently envisaged domain. The consultation of a broad range of experts is recommended to formulate an efficient preservation plan. Using a common vocabulary and reference model (such as the OAIS model) will enable improved clarity and understanding. One of the key goals of a long term archiving/preservation strategy is risk mitigation against loss and corruption.

► The preservation objectives of an archive must be defined and articulated in its archival policy. The policy should cater for the requirements of both data providers and future users (the so-called designated community).

[1] For references, see below.

► A good governance regime between data producers and archives needs to be established to ensure that the policy is correctly implemented in the foreseeable future.

► An archive should be able to evolve – an archivist should not be rigid in his or her thinking or preparations.

Principle 3: Be selective and decide what to archive and what to dispose of.

Archiving is an economic issue, as well as a technical challenge. Long term benefits are likely to be intangible, so it is advisable to concentrate on short and medium term benefits. Long term archiving will prove to be less challenging if the medium term actions are considered, prepared and undertaken well. The survival rate for data might be better if selected (less) material is archived well, rather than a vast amount of material being archived poorly.

► An archive should define the retention period envisaged for each individual dataset, product or feature group. The archive should also preserve the documentation that explains what has been selected to be included and why. This means that there needs to be an explanation of why certain aspects of a given dataset are more important in the shorter and/or longer term (collection policy).

► When thematic and reference data comes from different datasets / organisations it is better to archive them separately. More research is required in this area to identify best practice on how to combine them or archive them in such a way that they are in sync with each other.

► For data that is continuously updated, some consideration is needed of the best way to capture change. The following two main approaches are possible. More work is needed to understand the advantages and disadvantages and how both can be optimally combined:

- Archiving snapshots of the complete dataset. This method is easier to implement, but may produce redundancy, and therefore require more storage. It may also make it more difficult to synchronise different datasets at a later date
- Archiving of change only updates / logs. This method minimizes redundancy, leading to a reduction in storage required, but is not proven to be invariant against software and systems changes.

Principle 4: Consider preservation timeframes of 1, 10, 100 years.

Thinking in terms of these timeframes will help high level planning of archival systems by identifying the key considerations in the stages of the GI's lifecycle. As Principle 1 states, this thinking should start early.

► The 1-10-100 years terms are just suggestions and can be adapted as appropriate to reflect operational thinking.

1 year: operational safekeeping, focusing on short term needs, proprietary formats and specialist solutions may be appropriate.

10 years: a strategic, business archive, the focus should be on reusability and access of data. We refer to this as a Transitional Archive as it builds the bridge between shorter term data providers' needs and archival needs.

100 years: Long term archive aimed at preservation. Focus on robustness against data loss and corruption, ability to curate and migrate. Data preferably held in flat files, open format.

► Accept that you will need to move data between archives, with different technical solutions; and plan for this. Access to data in a 100 year archive may be achieved through accessing replicated data from a 10 year archive.

► The 100 year archive will often be in the custody of a National Archive. For the transitional archive, decide whether National Archives or National Mapping Agencies (NMA) is responsible.

The progression of data through the archival lifecycle may be illustrated as in the diagram below.

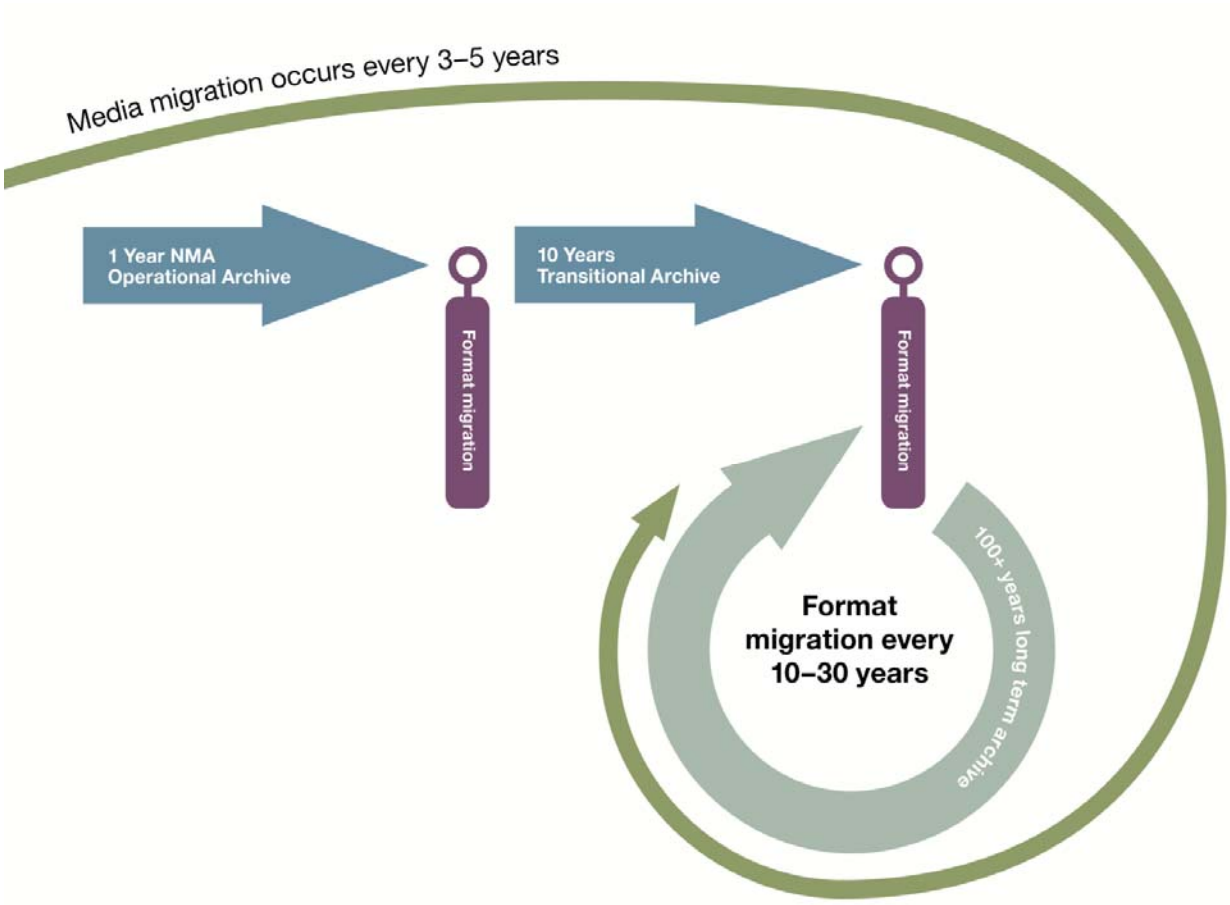

Figure 1: Archiving Lifecycle for Geographic Information

Principle 5: Migration or emulation is inevitable in the medium and long term. Be prepared and choose which properties to preserve in advance.

The encoding and format in which GI is handed over to an archive is often just one possible way of encoding the information and doesn't necessarily need to be preserved. It is highly likely that original data formats will not be supported and readable at some point in the future. Therefore, the capability to read all information and archive content needs to be retro-engineered (emulation) or the data has to be migrated into a supported format. Emulation is currently still in an infant stage. Therefore, we currently focus on migration as the main strategy to keep GI readable over a longer time period.

► Define the properties of your geo-information (i.e. significant properties) that must be retained at the point of migration or emulation. (Nestor, 2008 and 2011)

► Don't assume that archived data won't degrade - mechanisms for checking and reporting data degradation need to be put in place for all archives.

► Choose the appropriate data representation version (raw/processed, "products", originals/digital/analogue). Document your choice and rationale; this type of information will assist future archivists and users in understanding and determining the updated versioning requirements for the data.

► Enable, if possible, annotation of the archived data and preserve the annotation along with the associated contextual information to facilitate adding value to the data. Annotation should not change the original content of the data.

► Ensure amendments to archived data are recorded with an audit trail. Amended data should not pollute the authenticity of original archival data. A separate set of metadata to describe amendments and the audit trail may be necessary.

The requirement for an audit trail and addition of amendments is an example of why it is advisable to build in the capability for an archive to be able to evolve – an archivist should not be rigid in his or her thinking or

preparations. Collaborative automation may ease preservation efforts (see, for example, http://www.openplanetsfoundation.org/)

Principle 6: The output of the archival planning process should also be preserved over the long-term to accommodate future preservation requirements.

► Documents describing the archival planning process and policy need to be associated with the relevant geographic data in order to provide context for decisions made at the time of or before ingestion of data into an archive.

Principle 7: Archiving is not back-up. You must also back-up your archive.

► Backup is mitigation against catastrophic failure to enable Disaster Recovery and is considered to be a standard IT task in any operational system that holds business critical data. Archiving is aimed at the longer term retention and access to data information in a more managed way.

► It is necessary to back-up your archive on at least two uncorrelated storage systems. One backup system should be at a remote and secure site.

Principle 8: Geographical data should be preserved in a way that non geo-specialists can handle.

The likelihood that data survives and can be accessed will be higher if the data is structured in a way that archivists and users find familiar, such as other non-geospatial and mainstream content. A proprietary grid-format for an aerial photography image, for example, could be transformed to a format, such as TIFF, which archivists and users already use for a large amount of other archival image content.

► Document migrations, format, and structure so these can be understood by archivists, curators and users.

► Document the rationale behind applying certain preservation actions (e.g. migration) to the data. This type of information forms the preservation history of a dataset and will assist future archivists in understanding and determining the updated preservation requirements for that dataset.

► Independently[2] archive chronologically synchronized versions of thematic geo-data (e.g. climate or environmental data) and geo-referencing data (e.g. orthoimage or map data).

► Archive data specifications, definitions of coordinate systems and anecdotal material that will help to interpret and understand the data at a later point in time.

Principle 9: Information objects should be self-contained and independently understandable.

The more self-describing an information object is, the easier it will be to interpret it after a long period of time.

Principle 10: Keep the gold copy version of the 100 year data archive in open, file based repositories, not in databases, nor other complex environments.

An open repository enables more efficient data management, especially where machine-readable files are used, and supports future migrations. This approach also facilitates better, or at least simpler, access to the data, but be aware that there may also be some loss of information.

Principle 11: Consider keeping a graphical representation alongside the logical representation of the data.

When the logical representation of the data cannot be easily used by non-specialists, keeping a raster image or set of styled vectors of the entire geographical extent of the dataset for their use is recommended. When this is not feasible, keep an exemplary representation for a sample area to preserve how the data was typically rendered.

Principle 12: Restrict the number of formats and encodings to a widely agreed set of open, simple and well-documented file formats.

[2] See Principle 3.3

In the future, less effort will be required to either migrate or emulate data because fewer formats need to be supported. We expect the industry will come to a consensus about the most commonly used formats for preservation of GI.

► Support and adopt standards and best practice, e.g. processes.

► In the 100 year archive, do not use binary encodings with the possible exception for raster data. Any compression format for large text files or sub-packages (for example to bundle files that belong together in a container) should be open, well documented and widely used.

Principle 13: Prefer simple data models and schemas over complex ones.

A number of today's GI data models are rather complex, having been designed to support modern requirements such as to express rich attribution and relationships between entities.

► A highly normalised database representation, in which all properties of an object are spread around and need to be assembled for use, may be beneficial for short term access and data management purposes, but is not useful for long term preservation as there is a high risk that the linkage between the properties might be lost over time.

► While it may be desirable to preserve these, the balance between the preservation of the richness and the loss of data by having data models that may be very difficult to interpret in the future needs to be found on an individual basis.

► Consider, where feasible, migrating complex data into a simpler data model or structure at the point of ingestion to the 100 year archive. In certain circumstances (such as the format becoming obsolete) this may begin as part of the 10 year archival process.

Principle 14: Keep the access mechanism for archived data simple. Focus on basic current user requirements – an archival viewing system does not need to be a fully functioning GIS.

A simple mechanism and system that is easy to operate, maintain, migrate or emulate will be more beneficial in the long run than a system with lots of functionality.

► The long term, 100 year archive should only include basic access while more functionality can be provisioned to replicated data in a user environment based on short-term requirements.

► Decide what type of access service level is necessary based on the relevant context – static, stable on-site use or portable, platform-independent use.

Principle 15: Ensure effective management and quality assurance of the metadata associated with your data.

► Metadata stored in the archive should be both syntactically and semantically valid. For example, an XML-based metadata record can be validated against the corresponding XML schema to ensure structural validity. Semantic validation is more complex and may involve the use of controlled vocabulary defined by the archive, preferably through collaboration with the user community.

► Define the types of metadata needed to enable efficient discovery, accurate rendering, continued understanding and re-use (e.g. significant properties) and effective preservation of your data over the long-term.

► Use appropriate, widely-adopted metadata standards and formats.

► Apply appropriate and efficient versioning mechanisms to manage changes made to the metadata in the archive over time.

► Consider enabling the users to annotate the metadata in the archive to facilitate adding value to the metadata.

► Define a set of broad and high-level principles that form the guiding framework within which the metadata curation (management) can operate. The metadata curation policy would normally be a subsidiary policy of the archival data preservation policy statements. There should be clear reference to the archival and curation rules concerning legal and other related issues affecting the use and preservation of the data and metadata. All these elements should be governed by an overarching archival data policy.

Principle 16: Make some assumptions about future use, but don't be too restrictive.

We believe that it is better to make some assumptions on future use and explicitly document these rather than attempting to conceive an elaborate and intricately detailed forecast. Issuing some high level guiding statements will help to focus, rather than constrain preservation effort. How and in what way these assumptions are made is an individual judgement call between GI and archival experts.

Understand and document, if possible, the level of knowledge, technical expertise, and other related practices of the user community that may impact on the way the preserved data is discovered, accessed and used. An archive needs to monitor changes in the user community over time. Any substantial changes should be analysed and be addressed in due time by the preservation measures applied to the data in the archive.

► Identify the potential consumers (human, software application, etc.) to whom the preserved data object(s) will be beneficial in terms of accurate interpretation and proper utilisation, and then document these as assumptions.

Conclusion to the work

These 16 principles are the outcome of discussions between NMAs and archivists drawn from institutions and legislative contexts across Europe. The ease with which the group discussed and agreed these principles (during five meetings spread over two years) supports the view that there is a significant amount of commonality in national approaches and contexts to what are in many ways a unifying set of challenges. Perhaps the bigger challenge, though, is bringing interdisciplinary experts together from across (global) national boundaries to facilitate continued and further refinement of consensus on best practices, combined understanding of different perspectives and learning from wider contexts. Certainly, a collaborative approach is advocated by the group in preference to unilateral publication of archiving policies. In fact, given a key challenge is retaining a means to access information secured in digital file formats over the very long term, founding all geographic digital data archives on a similar set of collaborative principles and standards is probably the best way to guarantee that our treasured data remains accessible in the years to come.

Further Reading / References

General

- OAIS (2003): Open Archival Information System (OAIS) Reference Model, ISO 14721:2003 http://public.ccsds.org/publications/archive/650x0b1.pdf
- PREMIS Preservation Metadata: Implementation Strategies (Standard Homepage) http://www.loc.gov/standards/premis
- Sarah Higgins (2008): The DCC Curation Lifecycle Model, in: International Journal of Digital Curation, vol. 3 http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48
- Nestor (2006): Catalogue of Criteria for Trusted Digital Repositories http://www.nbn-resolving.de?urn:nbn:de:0008-2006060703
- Nestor (2008): Into the Archive. A guide for the information transfer to a digital repository. Draft for public comment. http://files.d-nb.de/nestor/materialien/nestor_mat_10_en.pdf
- Nestor (2011): Guidelines to digital preservation: process model and implementation (German, English translation forthcoming) urn:nbn:de:0008-2011101804, http://files.d-nb.de/nestor/materialien/nestor_mat_15.pdf

Geographic Data

- Archiving of Geodata (2010). A joint preliminary study by swisstopo and the Swiss Federal Archives http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/geodata/geoarchive.parsysrelated1.59693.downloadList.93958.DownloadFile.tmp/preliminarystudyarchivingofgeodata.pdf
- G. McGarva, S. Morris, G. Janée (2009) DPC Technology Watch Report on preserving geospatial data http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee
- G. Janée, J. Sweetkind-Singer, T. Moore (2009): Final Report of the National Geospatial Digital Archive (NGDA) and Federated Archive Cyberinfrastructure Testbed (FACIT) Projects, http://www.ngda.org/docs/ngda-final-report.pdf

Special Aspects

- Blue Ribbon Task Force on Sustainable Digital Preservation and Access http://brtf.sdsc.edu/
- Steve Morris (2010): Appraisal and Selection of Geospatial Data White Paper, Prepared for Library of Congress http://www.digitalpreservation.gov/meetings/documents/othermeetings/AppraisalSelection_whitepaper_final.pdf

Working Groups

- EuroSDR Data Archiving Working Group: http://eurosdr.net/archiving
- OGC Data Preservation Working Group http://www.opengeospatial.org/projects/groups/preservdwg