Beyond the brick, for the past in the future, you find the archive!

Karin Bredenberg, National Archives of Sweden <karin.bredenberg@riksarkivet.se>

Jaime Kaminski, University of Brighton <j.kaminski@brighton.ac.uk>

Abstract

The statement that XML is dead[1] is as wrong as celebrating Christmas on midsummer night's eve! At least in our opinion. Imagine making an archival soup based on international standards using XML, with one municipal archive, two regional archives, five national archives and the European Commission's eArchiving Building block thrown into the mix. This is what we are going to attempt, let us set the stage and take you through the recipe!

[1]https://developpaper.com/is-xml-dead/ is a starting point for getting more information regarding the statment.

Table of Contents

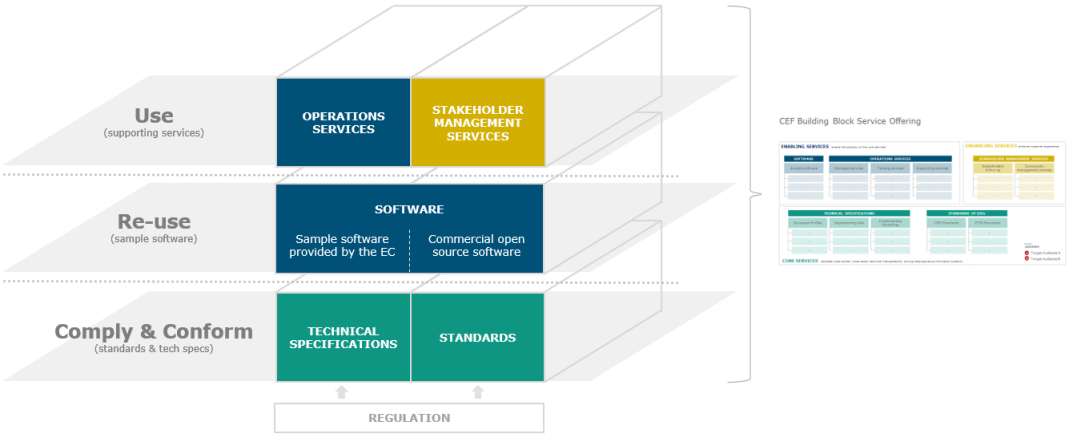
Archives

In our world the organisation responsible for saving the records of an organisational unit like documents or other artefacts that gives us our history is the archive. The archives exist in different levels of our society, in companies, in municipalities, in regions and at national levels as national archives. This means that the big difference between a library and an archive is seen in the library being responsible for the printed word. Today when the records are digital records the challenges to keep them has grown from taking care of paper to handling migration of records from obsolete formats, authenticity of the record so its reliable and most of all making sure the record is readable in the future considering format and available hardware.

Building Blocks

Let's start with Connecting Europe Facility[2] and its Building Blocks[3], what are they? The European Union realise that internet and digital technologies are transforming our world. A true statement. They also see is that the digital landscape is becoming more diverse, creating challenges for cross-border interoperability and intercommunication. Europe is about working together but, Europeans still face

barriers when using (cross-border) online tools and services. The implications are considerable. EU citizens can miss out on goods and services and businesses in the EU miss out on market potential, while also the different governments in EU cannot fully benefit from digital technologies. The EU has therefore described the Digital Single Market[4] (DSM) through which it aims to overcome these challenges by creating the right environment for digital networks and services to flourish. The DSM is not only achieved by setting the right regulatory conditions, but also by providing cross-border digital infrastructures and services. So, to support the DSM, the Connecting Europe Facility (CEF) programme is funding a set of generic and reusable Digital Service Infrastructures (DSI), also known as Building Blocks. The CEF building blocks offer basic capabilities that can be reused in any European project to facilitate the delivery of digital public services across borders and sectors. Currently, there are eight building blocks: Big Data Test Infrastructure, Context Broker, eArchiving, eDelivery, eID, eInvoicing, eSignature and eTranslation. The main part of CEF is a Core Service Platform, provided and maintained by the European Commission. Depending on the building block, the Core Service Platform may include technical specifications, sample software and supporting services (funding for the European Commission). The CEF building blocks offer basic capabilities that can be used in any European project to facilitate the delivery of digital public services across borders. The basis for the CEF Building Blocks are interoperability agreements between European Union member states. The aim of the Building Blocks is thus to ensure interoperability between IT systems so that citizens, businesses and administrations can benefit from seamless digital public services wherever they may be in Europe.



The Building block layers.

For each building block the European Commission provides a Core Service Platform which consists of three layers:

- At the core of each building block is a layer of technical specifications and standards that have to be complied with;
- To facilitate the implementation of the technical specifications and standards, a layer of sample software that complies with them and is meant for reuse (for certain building blocks only);
- To facilitate the adoption of the technical specifications and standards, a layer of services (e.g. conformance testing, help desks, onboarding services, etc.) meant for use (which varies depending on the Building Block).

All this means that the Building Blocks can be combined and used in projects in any domain or sector at European, national or local level.

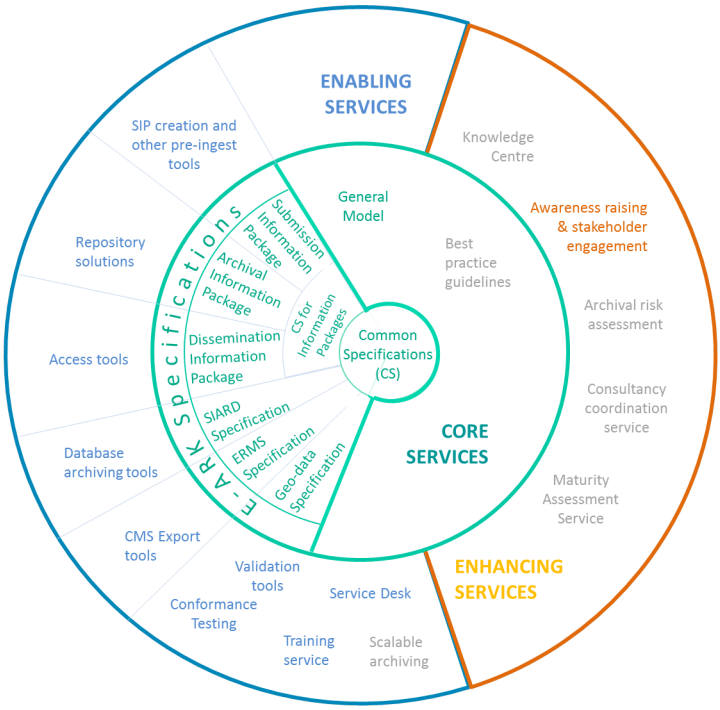
[2]https://ec.europa.eu/digital-single-market/en/connecting-europe-facility

[3]https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home

[4]https://ec.europa.eu/digital-single-market/en

eArchiving Building Block

For the archives and others transferring information the newly set up eArchiving building block [5]is the important component in protecting our history. The aim of eArchiving is to provide the core specifications, software, training and knowledge to help data creators, software developers and digital archives tackle the challenge of short, medium and long-term data management and reuse in a sustainable, authentic, cost-efficient, manageable and interoperable way. The core of eArchiving is formed by Information Package specifications which describe a common format for storing bulk data and metadata in a platform-independent, authentic and long-term understandable way. The specifications are ideal for migrating long-term valuable data between generations of information systems, transferring data to dedicated long-term repositories (i.e. digital archives), or preserving and reusing data over extended (and shorter) periods of time and generations of software systems. Next to the specifications eArchiving offers a set of sample software to demonstrate the format in different scenarios and business environments, and consultancy in regard to long-term digital preservation risks and their mitigation.



The eArchiving building block and its services and specifications.

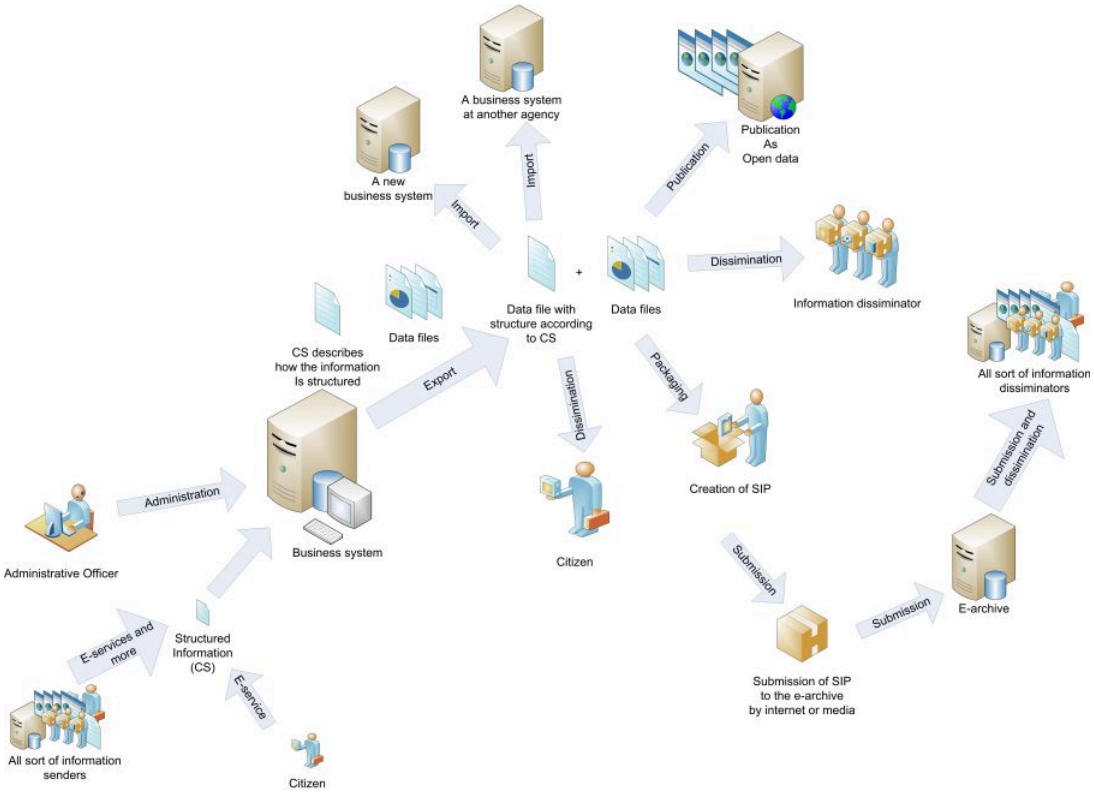
[5]https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving

Long-term preservation of information

Continuing from the specifications in the building block XML has long been the go-to format for the long-term preservation of information, mainly because it can structure information in a way that isunderstandable to both humans and machines. There have been concerns and comments raised regarding Json taking over the role of XML. However, in the archival setting even if JSon is easier to use as a programmer in the long term XML wins because readable elements and attributes explaining the information or what we call content placed in the document. However, at the same time XML needs to be used wisely with the correct element and attribute names to facilitate the understanding of the human reading of the XML-document both now as well as in the future. There are also other formats used like tiff and pdf but for reuse of the information XML is the format to go to in most cases.

Use cases

Our use case involves transferring information to different recipients. Information described with numerous ISO as well as *de facto* standards, all built around XML, which can be used for the transfer, storage and dissemination of information. Some of the most important use cases are shown in the following image.



The eArchiving use cases.

A short description of what we see in the image is that information is created in a system, the information is being exported as a number of different documents which can be sent to an electronic archive, being used by a researcher, published on-line, imported into a new system and so on.

Skill set

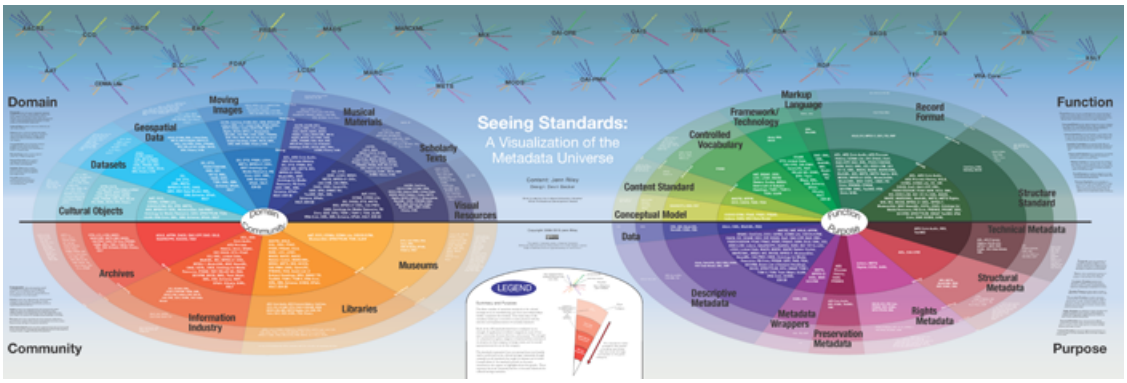
And with that setting of uses cases and need for reuse in a lot of ways we can only make the conclusion that a modern archivist, or perhaps the more accurate title is information or records manager, needs to understand XML and more than one standard for handling the information for which they are the caretakers. Besides the need for knowing a lot of different technical languages there is also a need to speak the same language as the programmers who will aid you with creating export and import tools. A challenge still not fully addressed in either community. The language challenge is something all the different communities and professions working together need to start working with. A common solution is that every work place and profession has their own language making cross-border professional exchange almost impossible without a lot of extra meetings, things that can be avoided if a common vocabulary was agreed upon and communicated to all professions. Look, for example, at the term ‘archive’ which means “to transfer records from the individual or office of creation to a repository authorised to appraise, preserve, and provide access to those records”[6] in the archival setting and the meaning “to store data offline.”[7] in the programmer and more technical setting. A unification is needed and at the same time not easy to achieve due to the different professions lack of common foras for these kinds of discussions (let’s not forget that not speaking to each other is more common than we would like to imagine).

[6]https://www2.archivists.org/glossary/terms/a/archive

[7]https://www2.archivists.org/glossary/terms/a/archive

Standards, de facto standards and specifications

Jenn Riley's Visualization of the Metadata Universe[8]has been around for a while, but still gives the best overview of standardsused in the cultural sector, archives, libraries and museums. In the image there are numerous standards are displayed in the context of their function and where they are used.



The visualization of the Metadata Universe by Jenn Riley. (The recommendation is to look on-line)

Almost all standards on the metadata map created by Jenn Riley have an XML format available described with an DTD or an XML-schema. For the XML-schemas both the ISO standard RelaxNG as well as W3C XML-schema formats are used. The choice depends solely on the skills of the creator of the schema and at the same time it's also common to ensure that all different type of schemas are available so transformations from RelaxNG to XML-schema and vice versa is often used. DTD are still around due to the fact that old software is still in use and they are based upon using a DTD.

The most common way to use the standards is to write a specification which describes a profile for our use case of the standard which then in its turn are implemented in the setting you are operating. The best way of describing the use of profiles is the standard METS (more about that later) which requires the user to write a profile describing how it is used.

[8]http://jennriley.com/metadatamap/

eArchiving Building block Specifications

The story of eArchiving Building Block specifications originates at work starting at the National Archives of Sweden and being enhanced in the E-ARK project[9]. The project delivered 7 draft specifications which all built upon creating profiles of standards and de-facto standards, these can be split into 4 specifications describing profiles of the standard METS and one for electronic records management, one for geodata and finally one for using the SIARD [10]standards. The project ended but no one wanted the work to be a usual project result, forgotten and not used so to make sure these specifications to not stop evolving the project created the "Digital Information LifeCycle Interoperability Standards Board" (DILCIS Board[11]). The board took over the specifications and have during the project E-ARK4ALL[12] (which are responsible for setting up and maintaining the eArchiving Building Block) brought them to a stabilised state and started the development of more specifications. A work carried out together with the experts of specific content so it's the experts writing the specifications. The board is set up with currently eight members and are charged with the task of handling the specifications. The Board is at the same time the core producers of specifications to the eArchiving Building Block.

[9]https://eark-project.com/

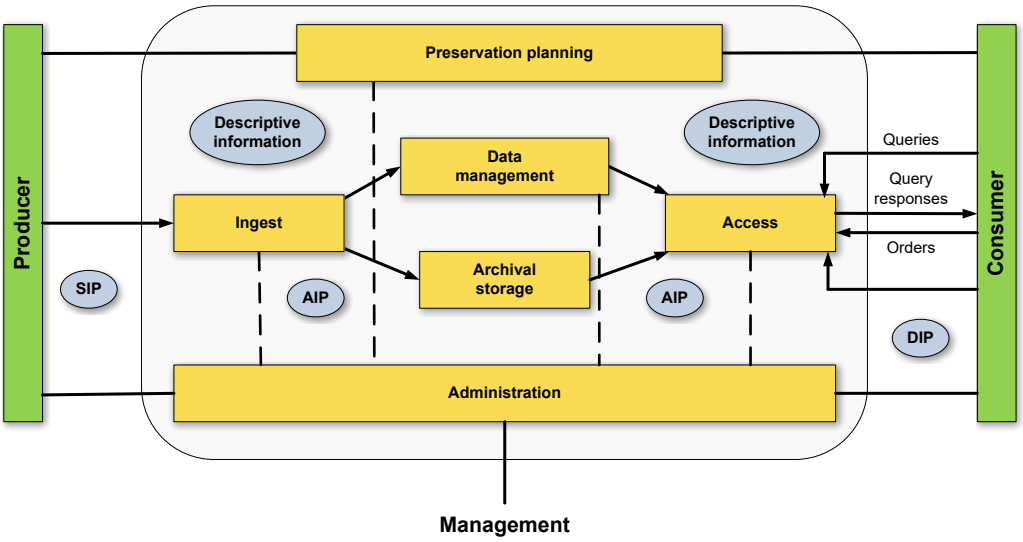
[10]https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

[11]https://dilcis.eu/

[12]https://e-ark4all.eu/

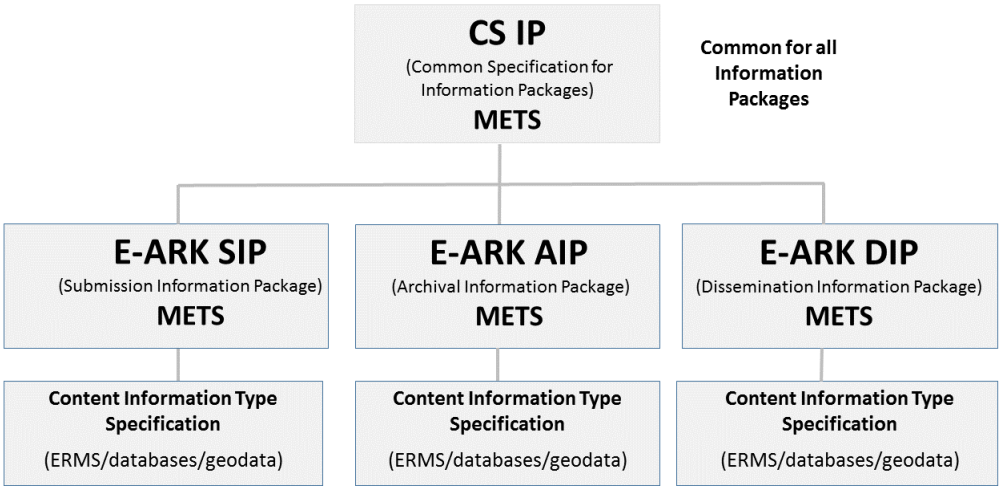
The eArchiving reference model setting

When we consider the European Commission's new eArchiving Building Block and that its evident that at its core are specifications based upon formats expressed in XML. The entireBuildingBlockrevolves around transfers of information where the final transfer is to the archives, but nothing prevents these transfers occurring earlier in the information lifecycle. The e-archive is built upon the Reference Model for an Open Archival Information System (OAIS Reference Model)[13]and its use of information packages, the Submission Information Package (SIP), the Archival Information Package (AIP) and the Dissemination Information Package (DIP). There are more parts described in the reference model but the core part used in the specifications are the information packages.



The OAIS reference model.

The different packages in the OAIS reference model are within the eArchiving Building Block described or get their inventory or manifest stated with the Metadata Encoding and Transmission Standard (METS)[14] which uses XML as the format for creating the readable text. (For the AIP it might be a format internal to the archival system you are using but that is another paper.) This way both machines and humans can understand the package (we do have some extra principles which aid with what constitutes a package). METS itself is a rather open standard with only one mandatory element being a structural description of the package. The standards also demand the creation of a profile for the exact use case describing the use of the standard and its elements and attributes. The profile for the eArchiving Building Block requires besides the METS XML-schema an extension XML-schema and validation rules in Schematron. The difficulties occur when the common user doesnot know how to validate the XML in combination with Schematron due to poor or no knowledge of either XML or Schematron and not having access to a person with the knowledge. We must not forget that Schematron adds its owncomplexities when it is combined with a non-relaxNG schema. A complexity that can be overwhelming when you don't know XML at all or just a basic understanding. This gives that numerousguidance documents need to be created ranging from how to write XML to how to use a specification. And all this since you cannot count on the person implementing the specification and its validation to have the appropriate background knowledge.



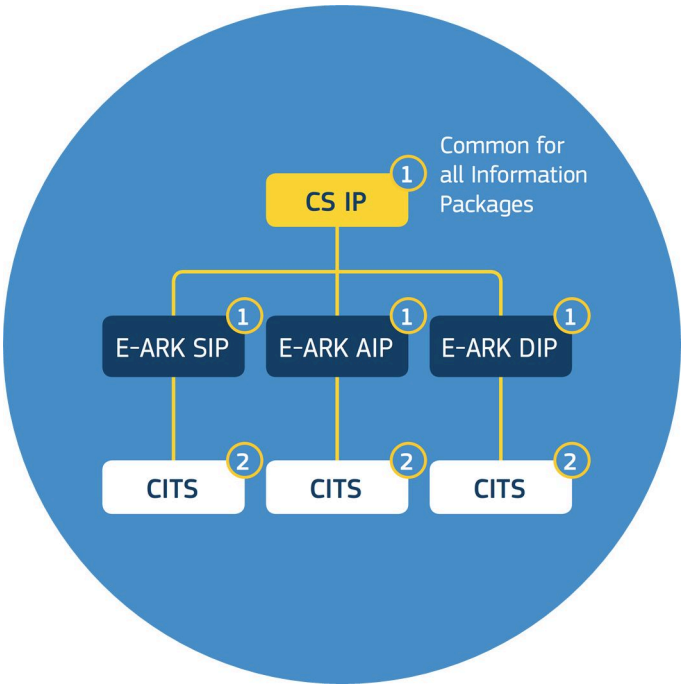
The eArchiving building block and the different specifications building up a Information Package. In the image the three different types of packages is seen.

[13]https://www.iso.org/standard/57284.html

[14]http://www.loc.gov/standards/mets/

Content specifications

Well, we now have a package described in XML, we need some content and maybe images, PDFs and information structured in XML that extends the information recorded in the package. This means we have one or more XML-documents to describe the information and its structure, and another XML-document to describe the whole package with all the content, but not the structure of the information itself. We can even handle a relational database in this way, extracting it in XML format and then packaging it in a SIP for transfer to the archive. Specifications of content are at the heart of the eArchiving Building Block, and the number of specifications is growing steadily. There are lots of different kinds of information that need to be described, luckily there are many specifications for describing information in XML, so there is no need to reinvent the wheel.



The eArchiving building block specifications.

In the image the content is described with the acronym CITS which means Content Information Type Specifications. Currently there are three available as described previously.

- A specification for electronic records management systems (ERMS), the specification uses a XML-schema as the format and is based upon several available records management standards which in their turn don't have a common XML-format available which means the ERMS specification is the connection between the different standards and the export of information from a ERMS.[15]

- A specification for geospatial data which uses the ISO standard for preservation of geospatial data in combination with the regulation within the union regarding geospatial data "the Inspire directive". In the specification a description of geospatial data and what it is found together with how the description of the data is carried out since the geospatial system themselves export information in readable formats but lack descriptions making the information understandable in the future.[16]

- A specification for how to place a relational database exported with the format SIARD in an information package. The format has been developed by the Swiss Federal Archives and is now a part of the DILCIS Boards responsibilities. The SIARD format is based upon export of the database as XML and several tools exist which aids with the task.[17]

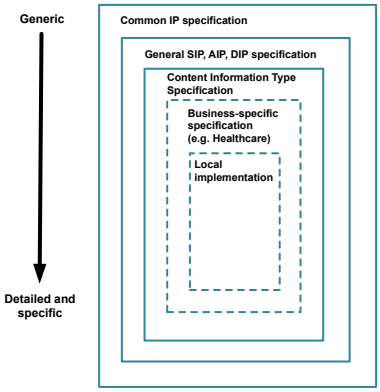
[15]https://github.com/DILCISBoard/E-ARK-ERMS

[16]https://github.com/DILCISBoard/E-ARK-Geodata

[17]https://github.com/DILCISBoard/SIARD

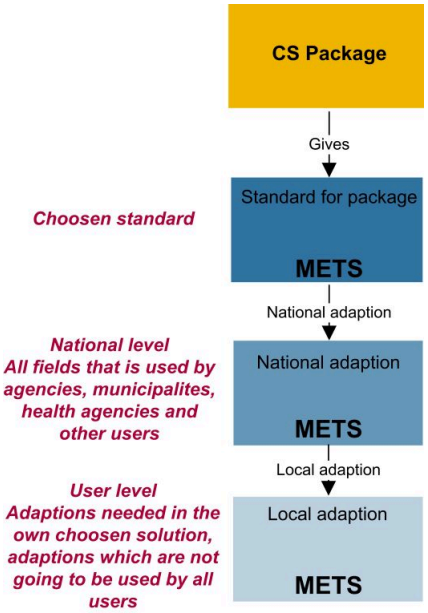
Basic soup recipe with a twist

We started with you imagine the making of an archival soup based on international standards using XML, with one municipal archive, two regional archives, five national archives and the European Commission's eArchiving Building block thrown into the mix. So how do we get this soup to be the "perfect" soup? We use core specifications based upon XML that can be enhanced the further down or closer to the user you get thus making the work and transfer to an archive a piece of cake.



The specification enhancement layers.

So our soup gets transformed into the set of use cases which is suitable for different kinds of users as well as different kinds of information.



The specification user layers of adoption.

The image shows us that the closer to the information and the implementation you get the more demands is added to make the information understandable in the future. We still are obliged to use the core set which means interoperability is achieved and at the same time the users own needs is an addition to enhance the core set.

Conclusion: Moving on from soup

The archival soup isn't in fact a soup it's a core set of common specifications using a machine and human readable language which makes it possible for one municipal archive, two regional archives, five national archives and the European Commission's eArchiving Building block to transfer information, store the information for "infinitive" time and deliver it to all different kind of users wanting to use the information, use the information for researchers, building applications, doing statistical reposts, prove

who they are, all the things you can do with information. This is supported by the advantages of XML and its way of being readable both by machine and humans. That the XML-document can be opened in just a text editor means that it is easy to read our past both now and in the future. This means that the support for XML needs to exist now and further down the road even if we in the archival community as XML users do not take part in all working groups maintaining XML instead being more concerned about the standards, *de facto*standards or specifications based on XML so we can ensure the past being created now will be present in the future.

Bibliography

[MDU] Jenn Riley: Seeing standards: A visualization of the Metadata Universe. Design: Devin Becker2009-2010, Work funded by the Indiana University Libraries White Professional Development Award. http://jennriley.com/metadatamap/