Centrality of data

Archivio di Stato di Lucca

New possibilities to bring the data closer to the owner

Dr. Giovanni Tartaglione

1. Introduction

This work contains the synthesis and motivation of the experience developed in Archivio di Stato di Lucca. In that Institute, the past experience produced a considerably quantity of data and of digital images, whose level is remarkable: among those more than 22.000 parchments, result of **IMAGO**, one of the most important project of Italian Cultural Heritage Ministry (MBAC). All that material, for many reasons, required a new intervention to let them live longer and be accessible to a number of users greater than in the past.
Moreover there were relevant documents (like historical maps) waiting from a long time for digitalisation to avoid the consumption due to the intensive access for the needs of architects and specialists.

In this occasion, even if the economic means were limited and mainly dedicated to digitalisation, the problem was addressed in its whole complexity: we wanted to make a virtual replica of all the Archive, what we call **e_ASLU**:

- to be filled gradually in the time and able to grow with the day by day work of Archivists, without the need of planning, every time, long lasting project;
- able to host old data without obliging to an hard work of normalization;
- directly accessible in Internet;
- able to handle images (also very large like maps can be);
- without Data Base
- with data written in clear form and used to be accessed in the native format;

The result of the project **e_ASLU** is accessible at the address www.archiviodistatoinlucca.it

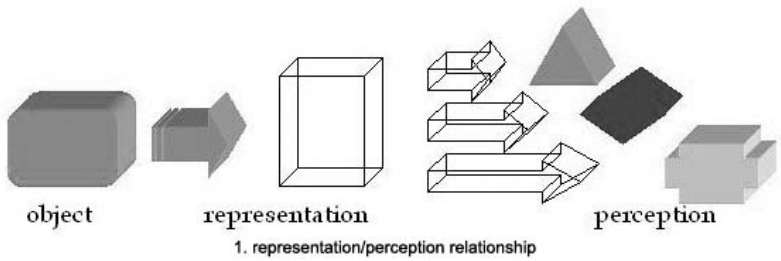
2. The patrimony

The property to be saved, in the hands of the Archive, is made of their own physical objects and of the data collected to represent those objects: the content.
We saw in the last years great interest about the relevance of "open source": it is an important argument that opens possibilities to the application developers but nothing to avoid software obsolescence and no change for data-owners whose attention goes to the object of their work: more than the application to collect their data, more than the perception offered to the user (that depends on the target addressed) , the content: the representation of the object made of its data and structure. The owner has the content in his focus; he needs to reduce expenses for his data management and for this, he requires long lasting, reusability, possibility of increasing, and wants to reduce as much as possible

the need to reformat the content according to new emerging normalizations for new and future applications.

3. Representation - Perception

The **Representation** of the object is the internal format of the object, containing the structure of the same and its characteristic data: it must be unique and not dependent on the single perception given to a target user class. The representation will not contain (as much as possible) elements of style and presentation.



1. representation/perception relationship

For the presentation of the object to the target, what we call the **Perception**, all data of the representation will be used in a way depending on characteristics of addressed user, on permissions, on aesthetic choice and so on. In is clear that for a single, unique representation there will be many possible perceptions and if on one side we would like to have a stable and long lasting representation, on the other, the perceptions will need to change because, as we say, times change.
So, what will be the relevant aspects of the representation, expected by the owner?

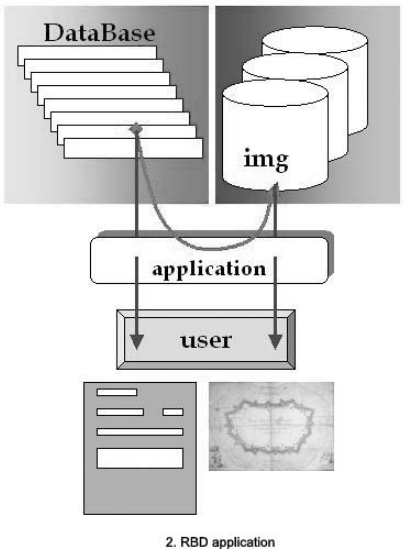
- *Long life*
  Production of data is very expensive: they contain the knowledge and professionality of the people who produced them. Every time we loose data we loose something more that is "culture".
- *Be natively simple, transparent and directly readable*
  like text format that is still the simplest format to save infos, that only requires a low level (free) text editor and avoids dependency from sw special licence.
- *Be able to grow in quantity and in structure*:
  for the Archivist the possibility of adding in the time, on a representation, new Sections, new Documents, new data or new structure sections produced by a different point of view (archivist, artistic, historic, etc.)
- *No dependency on proprietary sw*
  because sw get old very fast and your data risk to become unusable, unreadable.
- *Be very closed to the structure of the object*:
  the representation of the object as it is in the nature (in the Archive) and inside its context. So avoiding to put all parchments together in a table, another table for maps, another for essays while they are mixed on the shelf and distributed in different sections (fondi, serie, filze, etc.)
- *Reusable*
  to be used in new different applications, for events, conference, printings, specialistic applications.

4. Relational Data Base

Had a promising start, when the customer of an application required a specific result for the user, the application developer required data necessary for that target , data arranged in n-tuples inside a DB table: fast for cross search, with some limits:

- Object structure is not well represented because not so rigid like RDB requires;
- The schema is not expandable and difficult to be reused;
- Fields are positional: their meaning is given by the position in the record that must be known;
- Data dimensions are always defined and specially for fields like descriptions or notes there are strict limits;
- Fields or group of fields that for their nature are optional or can be repeated, are difficult to be implemented;
- Life is tied to the sw licence version;
- Difficult to mix in the same environment objects with different structure.
- Ties among data are inside the application and natively invisible.
- All these aspects can be exceed with a certain cost, but finally consider that the owner is obliged to use the intermediate sw even to simply read the data and this is a strong limit to the ownership of data.

Lets say also something about the structure of an archive based on RDB: it is usually made of three different parts, the tables of DB that contain one logical record for each object, with related data and reference to multimedia annexed like images; the images that are saved in a separate storage and the application that ties DB record (the object) to the images and shows the user the two section together: only through the application you are able to collect together all data related to an object: it is possible to distribute data and images in a more natural way, so that it will be simple and immediate for instance extract an object or a selection of objects (like fondo or serie etc.) with their complete data and images.



2. RBD application

There are many limits in the traditional approach and it is evident that technology has developed a lot both on the hardware to get faster machines and in the software to get powerful means, easier to be used, that can be approached without a deep specialization and knowledge. We have already in front of us simple tools that like biro in the age of fountain pen cannot be refused simply saying that they are tool for specialists.

5. & the represented object ?

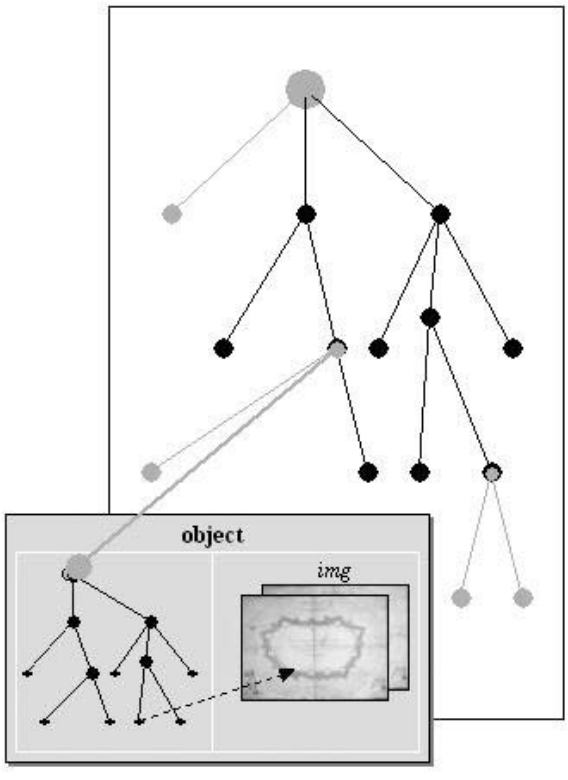
It is no more time of thinking to the single object or class of object completely abstract from the context: not simply the single charter, the map, the single document but the document, the charter, the map in the section inside the Archive and the same Archive inside a geographic distribution of Archives. There is the need of collecting information on

every level of the archive: the history of a collection makes more understandable the single document of the collection.
The target of our development was the image of the whole Archive with its heterogeneous content, its structure, its localization and description connected to history and culture of the Institute and its objects mixed as they are on the shelf (docs, maps, folders, …).

6. The solution

When we started, half 2006, for us it was time to implement a solution oriented to the representation of objects more than only data, taking care that in an Archive objects are not only the documents but also each level of collection of documents.
There was need to recover ancient data produced by past experiences without revision of their structure, with a representation coherent with new and future material without imposing rules of strict normalization that may change tomorrow, as it happened in the past.

It was necessary to take the archive very closed to the archivist, readable and simple to be modified, without the need of building very complex user interfaces for all maintenance activities.
It had to be written directly with the language and the format of the final run-time shape to be accessed by the user, avoiding intermediate passages and conversions.
The levels of the structure, corresponding to the different grouping of documents, are not always predictable: the concept of nesting solve the problem without obliging to a priori definition of max values.
This aspect is linked in some way to the capability of the structure to be expanded. We were thinking to the representation of all the Archive of Lucca but with the possibility to extend in a way that this archive could become a component of a father grouping (for instance all the archives of town, inside the collection of all towns of Tuscany and so on) to let the tree grow not only at leaves level but also at root level.



3. Tree structure

Two other performances like optional fields and fields that can be repeated, permit to better adapt the schema of a structure or of an object to the reality.

The structure that could better represent the needs of our vision is the **tree.** It contains data and object structure together with levels represented by nodes that have a simple backward reference to father and a self-explaining descendant structure.
A very long discussion took place with other application developers about the possibility of using a Markup or Meta-Markup language to describe in native mode the objects of an Archive. We saw a wide use of XML language and we were arguing that this could be the right language but also we found a strong opposition justified by the fact that this language has been used mainly to exchange data between applications. Moreover XML required a lot of memory and there was an intrinsic limitation on the dimension of XML files that over a certain amount become slow and unusable.
We must say that few people well know the characteristics of this language and its real usability as demonstrated by this project.
Often we have stated that with such Markup language, linear and readable, where the structure is defined in a clear way, the data and also the meaning of the data (that is all the knowledge of an object), the patrimony of the Institute, as defined at the beginning gets a longer life and a more wide usability.

Images of documents or collections are the heaviest part of the Archive. The project integrate with the data of documents also all their images that are often were very large or huge images. This part of the archive, that is the preponderant required the maximum simplification possible, avoiding to produce different series of images with different characteristics, detail, density depending on a specific use for each one. For this reason the decision was to use only one sequence of images, with maximum detail and quality, to be used for thumbnails up to 100% zoom to obtain the maximum detail possible. Limitations to the use of images could be imposed at visualization level, during the user access.

One of the aspects that we decided to leave unresolved is about the general question of maintenance: in primis the method to insert by the Archivist the description and data of documents and levels. At the beginning of this article we spoke of taking the patrimony of data as closed as possible to the Archivist. We thought and we are convinced that the effort to understand the structure of such a solution and the normal tools necessary to handle it, is much more light than the cost of an interface to permit the archivist to enter and manage data. It is no more time of complex tools reserved only to the specialist (informatics); the minimum training necessary to use programs from the commerce, for instance to built a document description file can be a first step to better understand projects and to built own projects gradually more complex.

7. e_ASLU

e_ASLU is the solution built in the Archivio di Stato di Lucca, that is easily applicable in other environments.
e_ASLU is a tree structure that represent all the Archive.
Levels are unlimited and each level brings its own data. Each node has its own information: for instance a folder with its description, its title, the images of the cover, etc. and its "sons" documents.
Each node is member of a node father and so recursively up to the root of the tree.

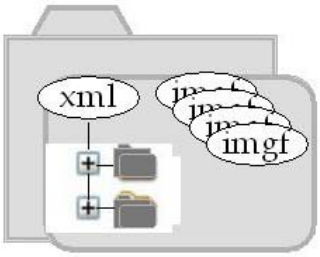
The Archive, that is a tree, can be transformed in branch of a tree (the collection of the archives or a region).
Leaf of a branch can be a collection, if no information is at disposal for the contained documents, otherwise documents can appear with their identification and content.
Leaves can be of different nature, (documents, maps, parchments, …) also mixed and have their own structure and data depending on the type and eventually on the shape of data already collected in the past.
Node structure reflects the nature of the level (fondo, serie, doc,…), contains a reduced number of obligatory fields, uses optional fields (like author), repetitive fields (e.g. title) and fields with extra specification (external, internal,…) according to the real shape of the node.

The tree contains structure and data inside; each node has references backward (ancestors) & forward (descendants). Each node is a natively complete object: the description of the node is stored in a /directory/ with its own images and this gets very clear the structure and gets very simple the selection of a node with its content (without the need of a specific application); the selection and extraction of a sub-tree is immediate.

As already said, the archive is built without the support of a Data Base, at least up to the functionalities of shipping in the tree. The technology used for the description of the tree and of all the contents is XML/XSLT to obtain all the performances described and expected. All the repository is natively written in XML.
Nodes and leaves representations are based on specific Schema descriptions that solve almost all the problems of coherence.
The perception of objects is built with XSLT files that permit to show according to the target-user selected, to the permissions associated to a profile, to parameters imposed by the Archive Authority to free or limit the access to information.

The logical tree of the whole Archive is translated in Repository on the disk with a structure of directories parallel to branches of the tree. Each node with related images is completely contained in a directory.
Each node is made of

- An XML file, descriptor of the node,
- More IMGF files with the images of the node,
- A subdir for each son



4. Tree item

With such a design, the generation of a new virtual tree to rearrange groups up to documents according to specific characteristics or parameters is very easy to implement, to give the user the possibility of shipping a personalized archive made of the only objects for his studies.

Images are stored in the Repository only once, in the most expanded format and max quality, and are accessed through an Image Server efficient and fast.

The Image Server XLIMAGE used in e_ASLU handles the original images in pyramid format for fast handling, without any need of thumbnail format, no secondary 72 dpi format. The image is shown in the visualization window, and can change very fast from all the image to max zoom of a detail in that window.

8. e_ASLU first content

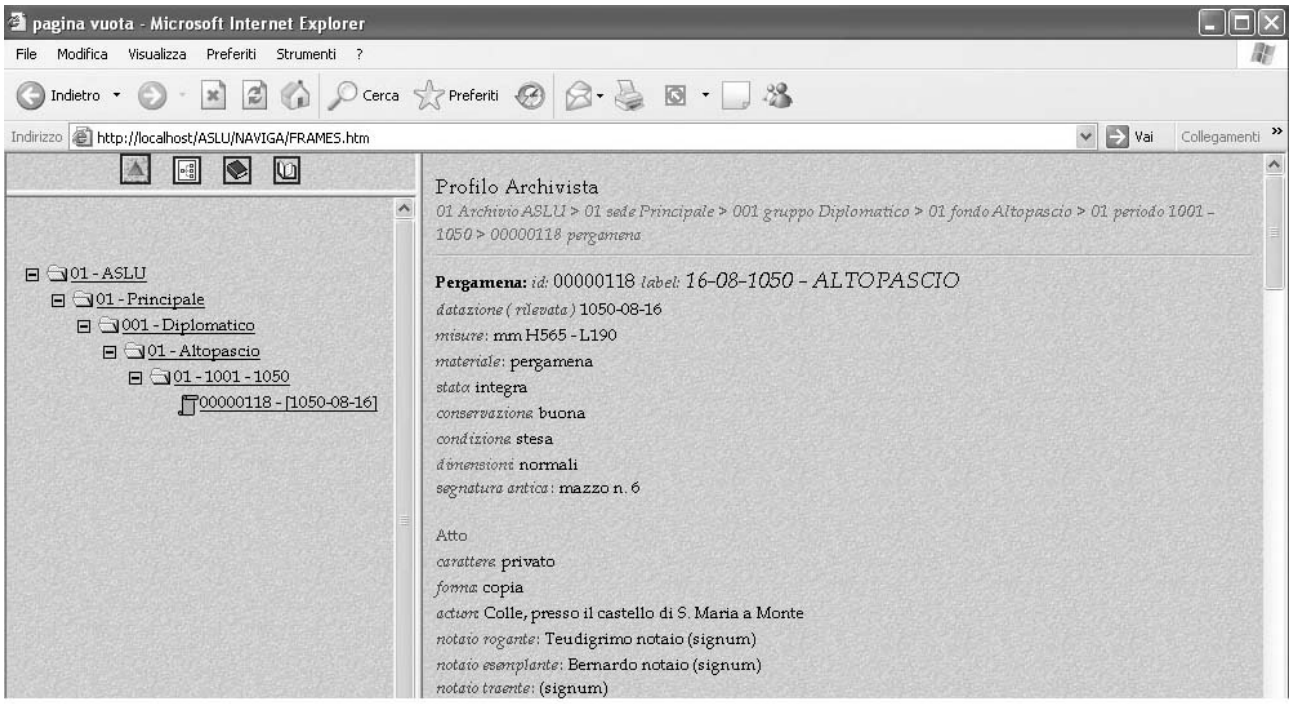
The first targets of e_ASLU project were limited to the

- Representation of the whole Archive structure;
- Filling the Diplomatic fund with 20.810 parchments and around 70.000 images;
- Filling Territory funds with more than 5.000 maps and documents and some 10.000 images;
- Shipping the tree, localization and visualization of content and images.

9. user interface

The user interface for the shipping and localization as clear as possible and consequently as easy as possible, to be understand and to be used.
When entering the web ASLU, Username and Passwords are requested to enter according to the profile with specific permissions given to the User by the Archive Direction.

After entering in e_ASLU branch, the window presented is shown divided in two parts:

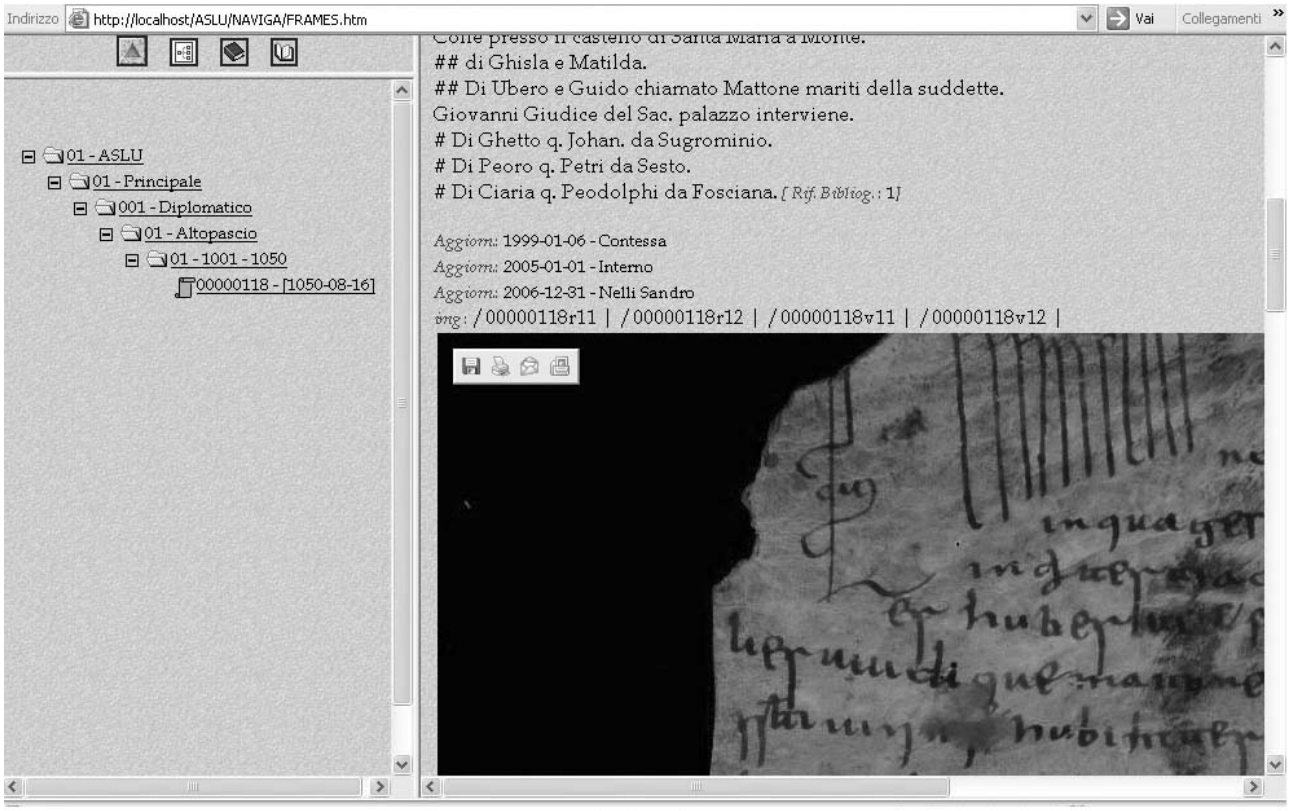


5, Shipping

- On the left**, Shipping** area **-** where the recursive structure of the tree is represented, with the same style of Windows file presentation, where each node is shown in the format,
  [+] Node_Name

- Click **[+]** to open descendant nodes
- Click **Node_Name** to show the node content on the right window

- On the right, **Content** area – where the node description and images are shown.



6. shipping

10.Indexes

The creation of indexes was not part of the current project, but some analysis has been taken to for the next step. Making indexes is one of the most expensive activity, for instance for a book, and especially if you want to index statements or sequences of words: in this last case you have to identify every occurrence of the sequence and flag it to permit the indexing function to create the right reference.
Previous projects developed also by the writer, put at disposal of the user a searching table normally reproducing the structure of the record, to permit complex searches on words present in different fields.
We have also verified that the specialist studying in the Archive in most cases don't use this complex method of searching but simply searches for a single word and then tries to reduce the selected objects with some further word, and so continues up to obtaining few results selectable by sight. It means that a good and sufficient result is reachable indexing single words, regardless the field where they are. This is a process that can be launched automatically without the need of man intervention, cheap and very effective; is the searching method at disposal on the most known search engines of Internet.

This will be argument of the next activities, and this kind or search will be adopted, together with an index of dates that could be helpful in time localization.

11. A strong question

There is in our Archives an argument often discussed that always divides Archivists, regarding the open access to information and images on Internet.
"What must be shown in Internet and what not?" "What must be published on Internet?"
We consider that today and tomorrow, in general for any organization, is very important to be present on Internet, and in a way that convince the arena that you are a focal point and a reference point for the area of activity in which you act, and this depends also on the **quantity** of material you put at disposal and on its high **quality** level.
So, possibly all material must be open and accessible and only if necessary, use intelligent limitations, if for instance the problem is to sell images: convince the potential buyer of the quality of data and of quality of images. The right policy could be to open completely data, excluding sensible information and let reach the maximum level of detail on images, eventually limiting the number of sights of the same image.

12. Conclusion

Some simple highlights to summarize conclusions:

- **e_ASLU**, Repository without DB licences, based on native XML/XSLT technology, and XL-IMAGE server.
- Increment of data time-life, economy of solution easy to be implemented.
- Possibility of fast increment and change of tree structure (add single node, add structure for new sight of the object
- Possible training of Archivists to let them to manage and take part on the informatics, avoid expenses to build complex interfaces, growth of personnel, increment of knowledge diffusion