Rationale and some principles for a VLDL data model

Nicolas Spyratos[1] Carlo Meghini[2]

[1] UP Sud - LRI, Paris
[2] CNR - ISTI, Pisa

Outline

1 Motivation and goal

2 Mathematical preliminaries

3 Digital Objects

- View
- Content

4 Relationship with the Web architecture

5 Conclusions

Outline

1. Motivation and goal

2. Mathematical preliminaries

3. Digital Objects
   - View
   - Content

4. Relationship with the Web architecture

5. Conclusions

Outline

1. Motivation and goal

2. Mathematical preliminaries

3. Digital Objects
   - View
   - Content

4. Relationship with the Web architecture

5. Conclusions

Outline

1 Motivation and goal

2 Mathematical preliminaries

3 Digital Objects

- View
- Content

4 Relationship with the Web architecture

5 Conclusions

Outline

1 Motivation and goal

2 Mathematical preliminaries

3 Digital Objects
   - View
   - Content

4 Relationship with the Web architecture

5 Conclusions

Motivations

Motivations:

- To facilitate the creation of Digital Libraries and the discovery, access and re-use of the digital objects in Digital Libraries (DLs).
- To create a yardstick, against which to "measure" DLs.
- To highlight the mathematical backbone of a DL.
- *As simple as possible, but not simpler.*
- Compliant with the Web (the largest DL ever, so far).

Goal

We need a level of abstraction over the overwhelming amount of details involved in the management of a DL, *i.e.*, a *data model.*

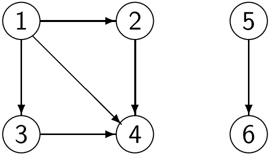
Operations provided by the model:

- *describe* an object of interest according to the vocabulary of the community;
- *discover* objects of interest based on content and/or description;
- *view* the content of a discovered object;
- *identify* an object of interest, in the sense of assigning to it an identity;
- *re-use* objects in a different context.

We want to define all concepts related to DLs and to give algorithms for the creation and management of a DL.

Mathematical preliminaries

We use directed graphs, represented as set-valued functions, as our basic modelling tool.



a	$f(a)$
1	{2,3,4}
2	{4}
3	{4}
4	{}
5	{5,6}
6	{6}

A : any non-empty set.

$\mathcal{P}(A)$: the powerset of A.

A *set-valued function* f on A is a partial function assigning to each element a in its domain of definition, a possibly empty subset of A :

$$f: \; A \longrightarrow \mathcal{P}(A)$$

$def(f)$: the domain of definition of f.

For each $a \in def(f)$, $f(a)$ is called the *image* of a under f.

$range(f) = \bigcup\{f(a) \mid a \in def(f)\}$

f partitions A into two subsets:

- the *active* objects, $\mathrm{act}(f)$, the objects that appear in f (either in the domain or the range of f):

$$\mathrm{act}(f) = \mathit{def}(f) \cup \mathit{range}(f)$$

- the *inactive* objects, $\mathrm{inact}(f)$, the objects that do not appear in f

$$\mathrm{inact}(f) = A \setminus \mathrm{act}(f)$$

a	$f(a)$
1	{2,3,4}
2	{4}
3	{4}
4	{}
5	{5,6}
6	{6}

$$A = \{1,2,3,4,5,6,7,8\}$$
$$\mathit{def}(f) = \{1,2,3,4,5,6\}$$
$$\mathit{range}(f) = \{2,3,4,5,6\}$$
$$\mathrm{act}(f) = \{1,2,3,4,5,6\}$$
$$\mathrm{inact}(f) = \{7,8\}$$

An active object a is:

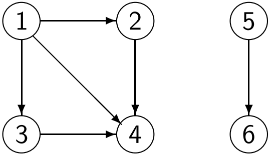
- *initial* if it is not in the image of any other object:

$$a \in \mathit{def}(f) \text{ and } [\ (\forall x \in \mathit{def}(f))\ a \in f(x) \rightarrow x = a\ ]$$

- *terminal* if either it is not an identifier, or it is an identifier and belongs to its own image:

$$a \in \mathit{range}(f) \text{ and } [\ a \in \mathit{def}(f) \rightarrow a \in f(a)\ ]$$

- *intermediate* if it is neither initial nor final.



initial: $\{1, 5\}$

terminal: $\{4, 6\}$

intermediate: $\{2, 3\}$

Digital Objects

A DL includes a set of digital objects.

A DL is very different from a traditional information system, which contains *representations.*

Intuitively, we think of a digital object as a piece of information in digital form such as a PDF document, a JPEG image, a URI and so on.

As such, a digital object can be processed by a computer, for instance it can be stored in memory and displayed on a screen.

More formally, let O stand for a collection of (digital) objects.

We assume O to be non-empty and countable.

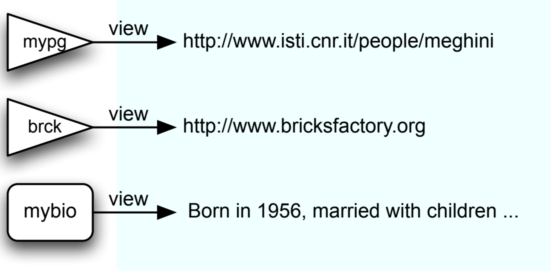
Objects have a *view,* a *content* and a *description.*

View

We assume that each digital object can be *viewed* using an appropriate mechanism.

Given $o \in \mathrm{O}$, $\mathrm{view}(o)$ denotes the result of viewing o.

view is a total function having the set O as domain.


mypg
view
http://www.isti.cnr.it/people/meghini
brck
view
http://www.bricksfactory.org
mybio
view
Born in 1956, married with children ...

Content

We define *content* over O to be a set-valued function cont on O :

$$\mathrm{cont} : \mathrm{O} \rightarrow \mathcal{P}(\mathrm{O})$$

such that for each object $o \in def(\mathrm{cont})$, the image of o under cont, $\mathrm{cont}(o)$, is a finite, possibly empty set of objects that we shall call the *content* of o. We shall call each object o in $def(\mathrm{cont})$ an *identifier*.

Starting with a content $\mathrm{cont}(o)$, one can obtain several different *documents*, each document corresponding to a rendering of $\mathrm{cont}(o)$ on a specific device.

Note: we do not exclude the case in which $o \in \mathrm{cont}(o)$.

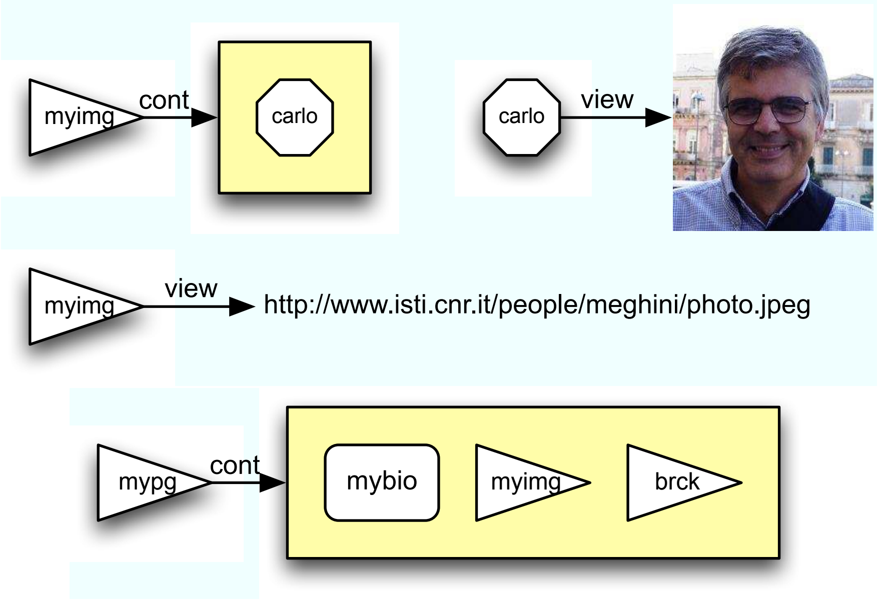
Note: content is dynamic (in time and space).

Inactive objects are not used currently in the library, but nevertheless are available to the user community. They may enter the content function either as identifiers or as elements of content at any later point in time.

Initial objects: identifiers of *collections*.

A special category of initial objects: objects with empty content, *i.e.* such that $\text{cont}(o) = \emptyset$.

A terminal object: a “pure” content object, in that it is not associated to any object. A terminal object contributes to the DL by its visualization only, *i.e.*, by applying to it the function view.



myimg cont carlo

carlo view

myimg view http://www.isti.cnr.it/people/meghini/photo.jpeg

mypg cont mybio myimg brck

Relationship with the Web architecture

The web architecture is based on three fundamental notions: *resource, representation* and *identifier.*

- A resource “can be anything that has identity”. An *information resource,* in particular, is a resource all of whose “essential characteristics can be conveyed in a message”.
- A representation is “data that encodes information about resource state”.
- An identifier is “an object that can act as a reference to something that has identity”. To achieve global communication, “the Web makes use of a single global identification system: the Uniform Resource Identifiers (URI)”.

$$\begin{tikzcd}
\mathcal{P}(O) & & \text{id} \arrow[ll, "\text{cont}"'] \\
\text{representation} \arrow[u, "h"] \arrow[r, "\text{render}"] & \text{inf. resource} \arrow[ul, "g"'] & \text{URI} \arrow[u, "k"'] \arrow[l, "\text{deref}"']
\end{tikzcd}$$

- h associates each representation to the set of objects it contains
- k associates each URI to an identifier, 1:1
- g associates each resource to the set of objects it contains, so that:
  1. for each representation r, the digital objects in r are those contained in any rendition of r : $g \circ \text{render} = h$
  2. for each URI u, the digital objects contained in a resource identified by u, are the content of the identified corresponding to u : $cont \circ k = g \circ \text{deref}$

We can prove that, given two functions h and k satisfying the conditions above, there is a unique function g which satisfies the above.

Conclusions

The beginning of a model of a DL, compliant with the web architecture.

Thank you!

Any question?