SWIB 2012:
Workshop on Metadata Provenance

Part 4:
Case Study: DM2E

Digitized Manuscripts to Europeana

(Putting it all together)

Agenda

Review Europeana Data Model

OAI-ORE vs. Named Graphs

Linked Data Publishing with Provenance

Status Quo Europeana

Europeana provides data about **cultural heritage objects** (**CHO**) from CH institutions all over Europe.

Provenance requirement: Distinguish metadata from different institutions talking about the same (owl:sameAs) resource.

DM2E Mission Statement

DM2E develops infrastructure to:

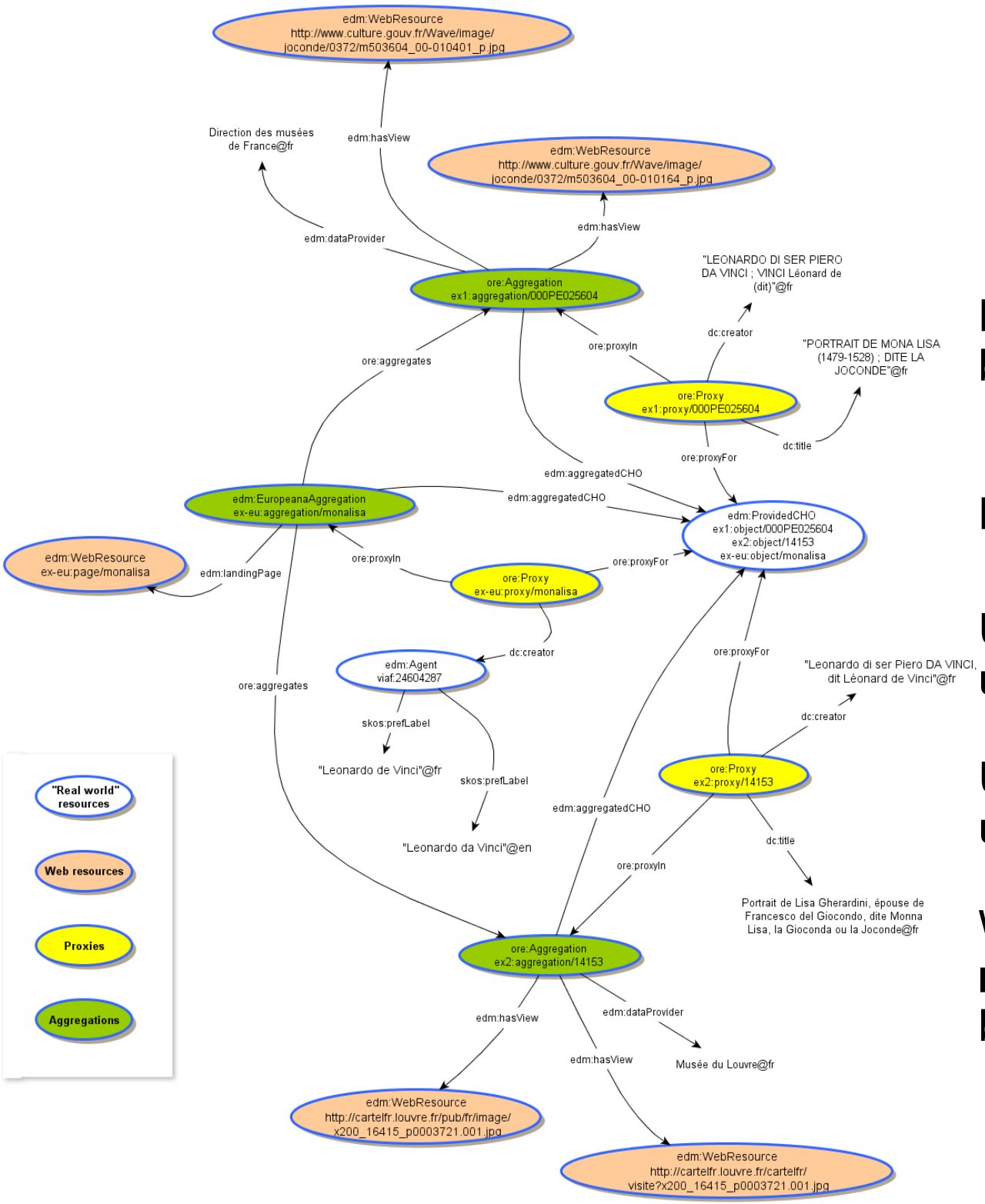
1) ingest metadata from data providers,

2) transform metadata to RDF,

3) provide RDF data for consumption by researchers from digital humanities,

4) deliver RDF data to Europeana.

The Europeana Data Model (EDM)



Provenance realized by means of OAI-ORE.

Problems?

Users have to understand **Proxies**.

Users have to understand **Aggregations**.

Wouldn't **named graphs** be nicer?

How are proxies and aggregations used?

What is an aggregation?

> *"Aggregations are used in Europeana to **represent the complex constructs** that are provided by contributors. An aggregation **is associated to the object that it is about**, by the property edm:aggregatedCHO."*

Level of aggregation:

1 aggregation per providedCHO.

EuropeanaAggregation aggregates other aggregations (from data providers).

Removing the proxies

Proxies are (proxy-) resources for the actual resources. Every data provider has an ″own″ resource to describe, as a **placeholder**.

But: Data providers use different URIs for their resources anyway.

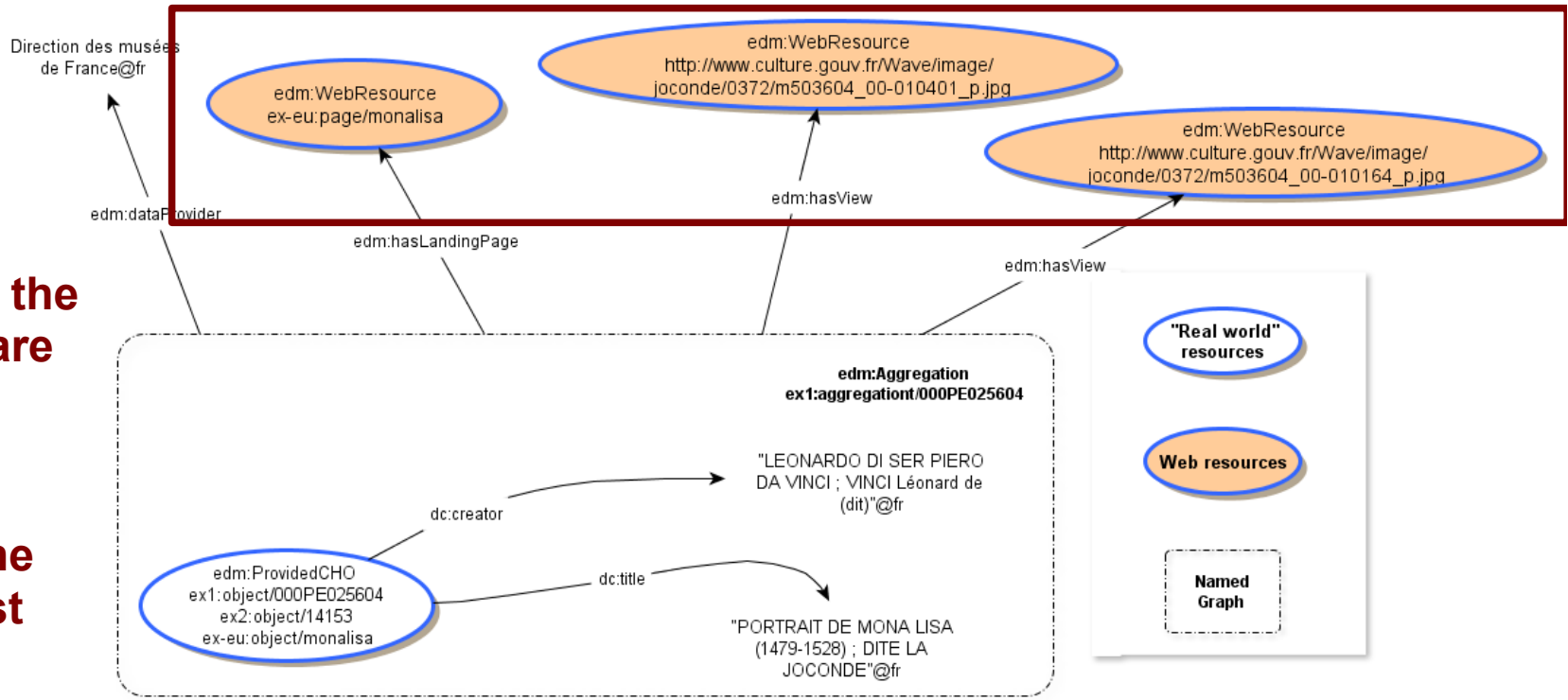
Linking creates owl:sameAs statements and conflates resources.

How can we then reliably maintain different descriptions?

We simply use **named graphs** to distinguish descriptions from **different providers**.

A Named Graph per Resource

Corresponds to the EDM aggregations.
Finegrained… feasible?
Named Graphs as first class members in the model.



Statements about the aggregation that are only valid for one resource!

If we allow this, the named graph must never get lost!

Nested Graph Problem

Named Graphs are connected to Linked Data Principles.

One Named Graph per document fetched from a URI.

If we provide a dump of the full dataset from one provider, we have several named graphs within one resource that form another named graph: **Problem!**

This is the (still unsoved) Nested Graph Problem.

RDF will not provide a solution, it is not clear how to deal with such data.

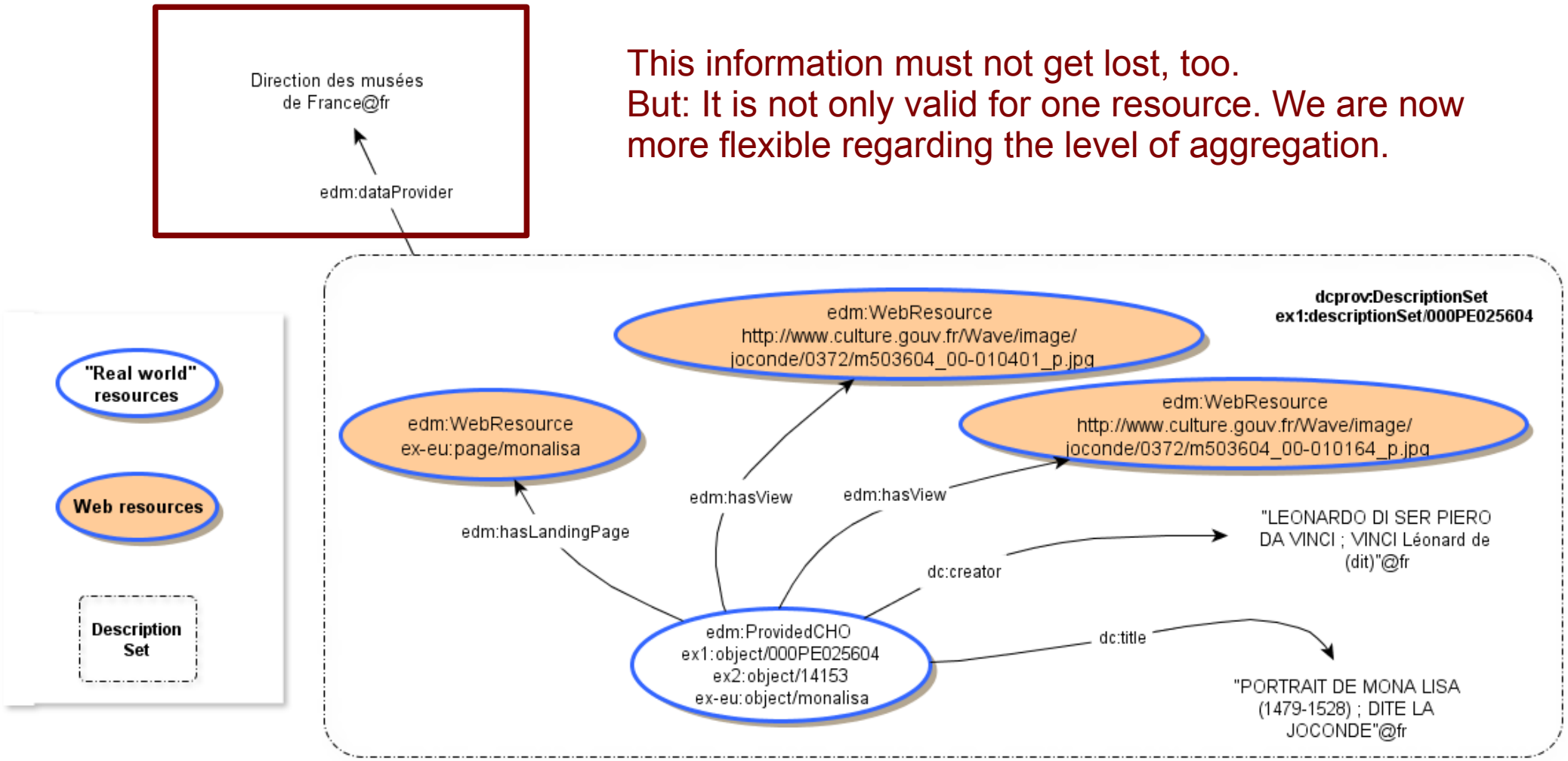
Linked Data Provenance Mantra:

Do not publish Named Graphs!

Because everything you publish, IS a named graph.

But use them to manage your provenance.

Solution: A Named Graph per Provider



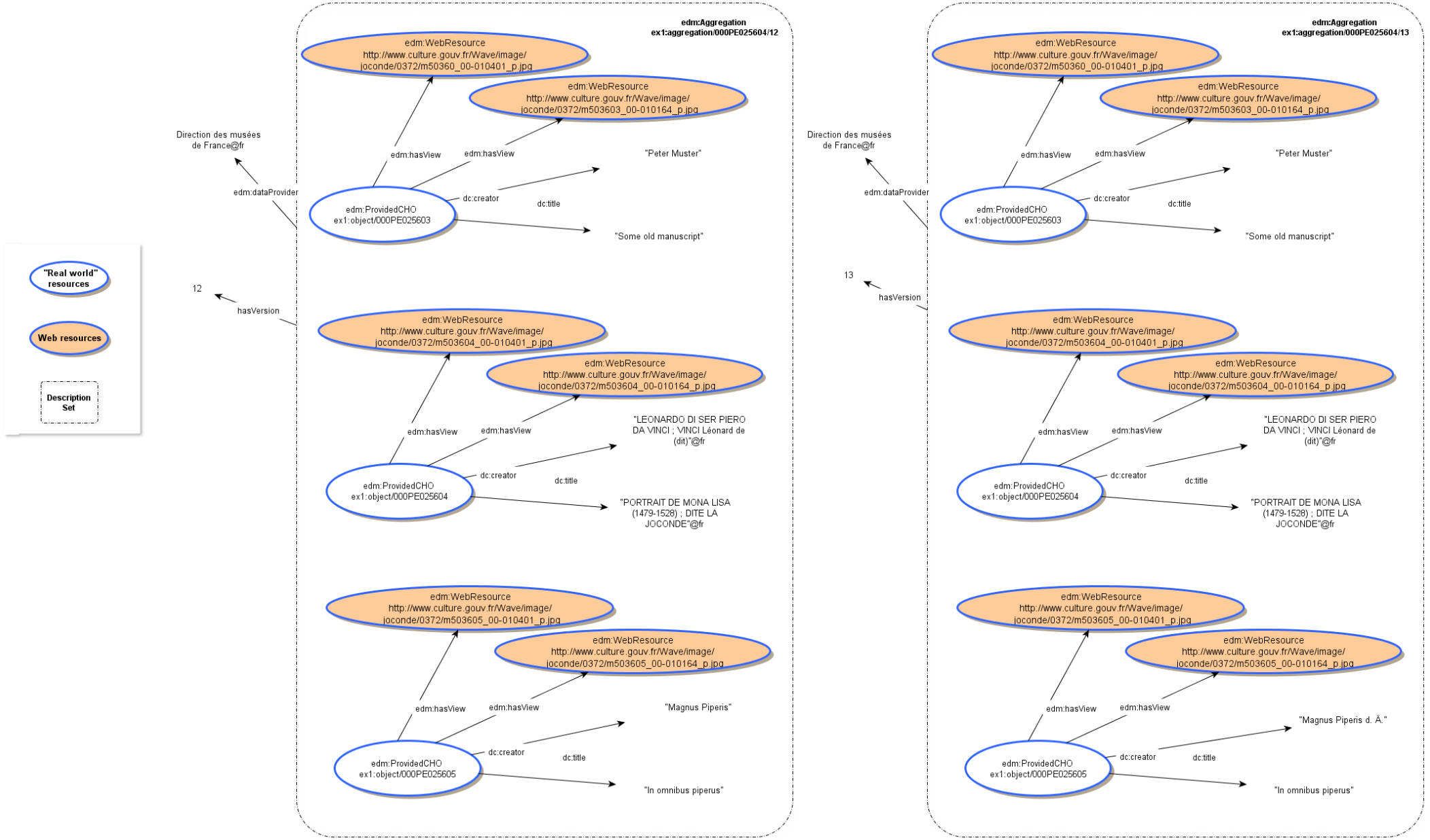
This information must not get lost, too.
But: It is not only valid for one resource. We are now more flexible regarding the level of aggregation.

One Named Graph per Provided Dataset

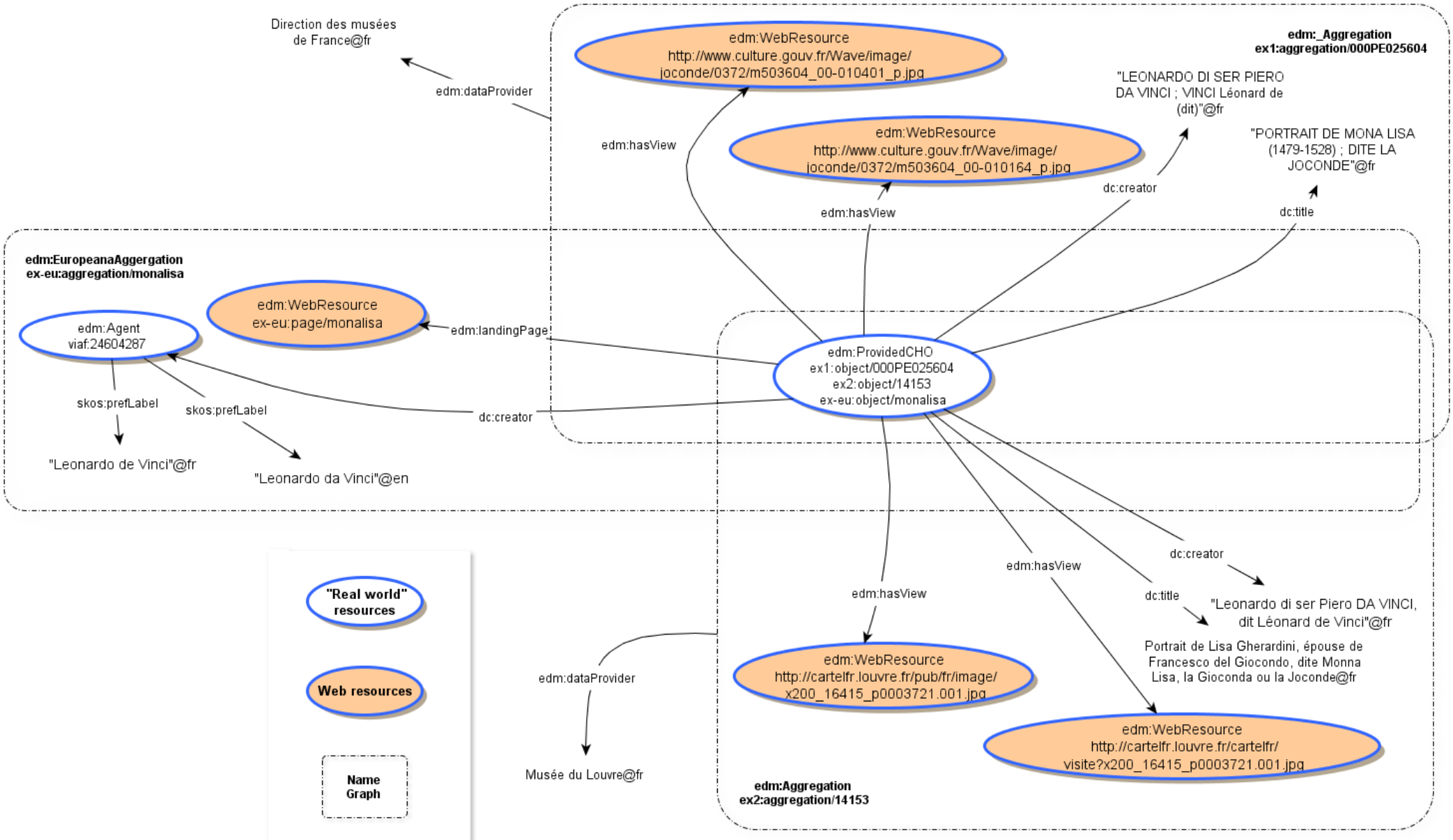
Naturally fits to provenance requirements:
All statements stem from some dataset.

Positive aspect: Dataproviders do not have to **care any more!**

Provenance and Versioning on Dataset Level



Overlapping Resource Descriptions



Crosswalk to EDM

Are we still backwards comaptible?

1) Create one Proxy per providedCHO.

2) Create one Aggregation per providedCHO.

3) Assign all statements about the Named Graph to the Aggregation.

4) Assign statements about the resource to the Aggregation that were originally Aggregation statements (landing page, web views).

5) Assign all further statements about the resource to the Proxy.

6) Assign Proxy and Aggregation to the providedCHO.

Publishing

Remember the Mantra:

Do not publish Named Graphs!

What's inside our store?

RDF Datasets, organized in named graphs.

NG URI scheme:

http://data.dm2e.eu/data/dataset/[provider]/[datasetId]/[version]

Additional provenance statements for each graph.

Make it available

Web-Documents (with URI) deliver RDF, provenance is included as statements about the URI.

On client side, the document creates a new Named Graph, with the URI as name.

RESTful API (Publishing)

http://data.dm2e.eu/data/…

… **dataset**/[provider]/[datasetID]/**[version]**
=> dump of one whole ingested dataset

… **resource**/[provider]/[identifier]
=> 303 to latest version

Can we have more than one dataset from one provider talking about the same resource?

… **dataset**/[provider]/[datasetID]/**[version]**/[identifier]
=> data about a single resource

… **linkset**/[DM2E]/[linksetID]/**[version]**
=> generated links

… **linkset**/[DM2E]/[linksetID]/**[version]**/[provider]/[identifier]
=> links for a specific resource

Hint: Documents contain a **[version]**.

Provenance in Documents

Generated from provenance information about datasets:

- dc:creator => Data provider
- dc:date => Timestamp
- dm2e:version => version number
- dm2e:nextVersion => link to next version of the document
- dm2e:previousVersion => link to previous version
- dm2e:links => link to a linkset
- Optional: PROV statements for full provenance chain.

Maintained by the DM2E infrastructure.

Version means always the version of the underlying dataset.

Consequences

1) Besides new versions, there are no datasets from one provider talking about the same resource (Persons?)

2) No named graphs on "payload" level (EDM+), there is no such thing as an ore:Aggregation per resource any more.

3) …

Implementation pending ;-)

Questions?

Suggestions?