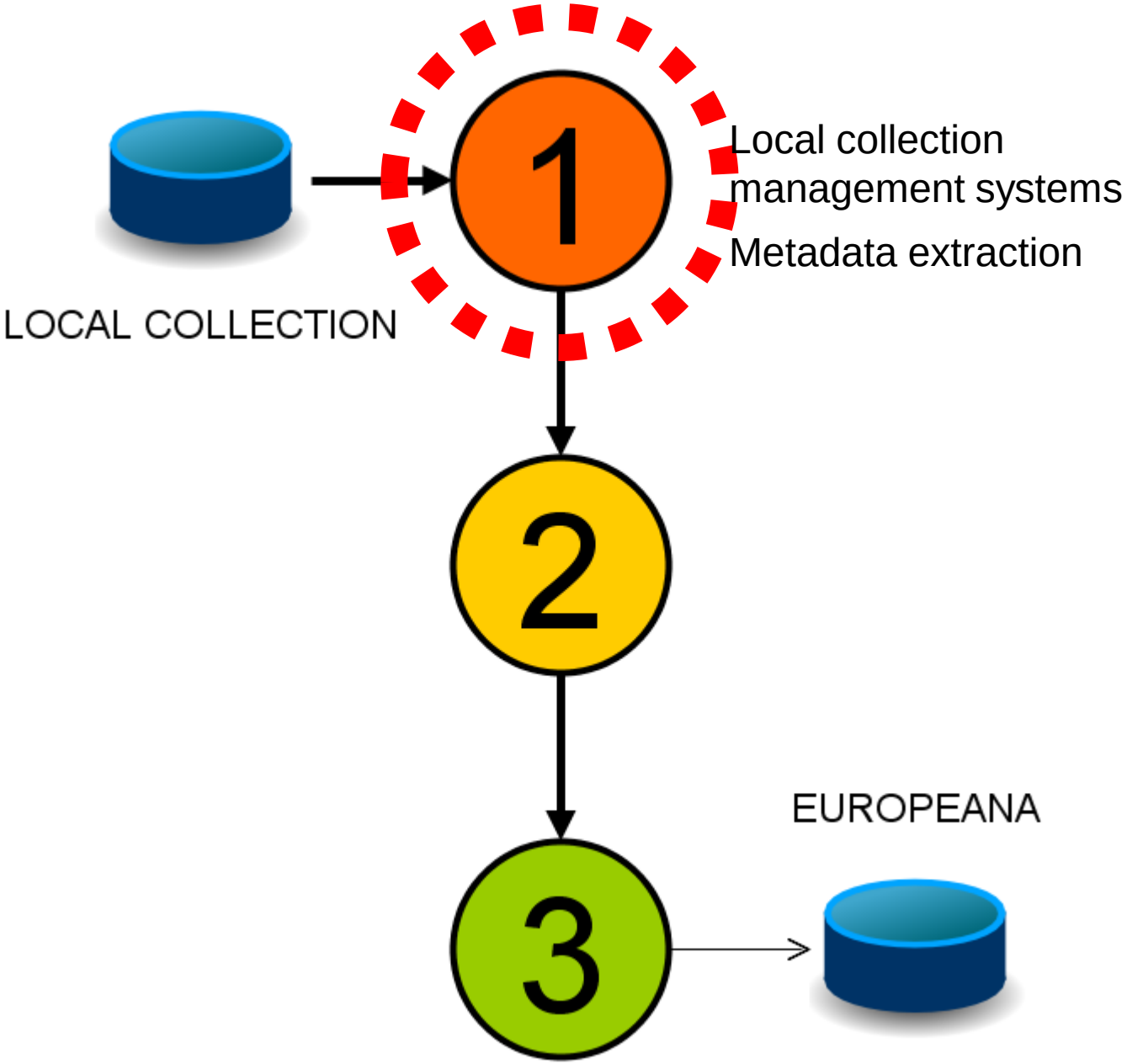
CARARE Training Workshops

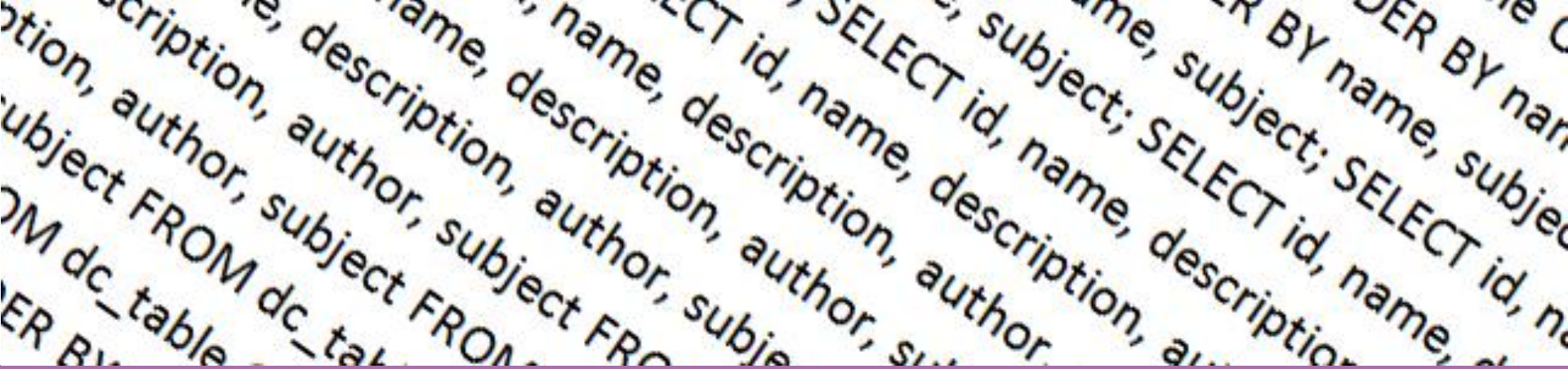
Stein Runar Bergheim
Asplan Viak Internet as

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Metadata Transformation





Metadata Extraction

Extracting data from databases into XML

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

We need to get data from your database into XML

The first step on our way is to export our data from our database to XML

For this reason – we need to have some base knowledge of both databases and XML

What is provided here is overly simplified – but quite possibly sufficient to do what needs to be done

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

```
<?xml version="1.0" encoding="utf-8"?>
<root>
        <element attribute="value">
                Text
        </element>
</root>
```

Example XML document

What you need to know about databases

Information stored as tables

Each table defines a set of fields

Each table has rows with values corresponding to the fields

Tables have identifiers

- Primary keys – unique ID for the table row
- Foreign keys – an ID referring to a value in another table

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Creating XML from your database

Exporting via native interface (built-in function in collection management system)

Writing custom export script

- Involves a query language (typically SQL)
- Involves a scripting language
  - VB/C#.NET
  - PHP
  - PERL
  - Others

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

XML is picky

Escape special XML characters (<, >, &)

Convert to UTF-8 version of Unicode

Convert entity references (e.g., ©)

Remove extraneous whitespace

URLs

- /?#=&;+ must be encoded as “escape” sequences

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Splitting and merging

In the source database:

Location (free-text field):
"...Place: London
North: 51 3′
East: 0 7′..."

In target format:

<geoName>London</geoName>
<geoPosition>51.3, 0.7</geoPosition>

Or:

<location type="geoName">London</location>
<location srs="epsg:4326" type="geoLocation">51.3, 0.7</location>

Not:

<location>London</location>
<Location>51.3, 0.7</location>

When extracting data we need to consider which information can exist in text fields and which information needs to be split into atomic components.

Metadata Normalization

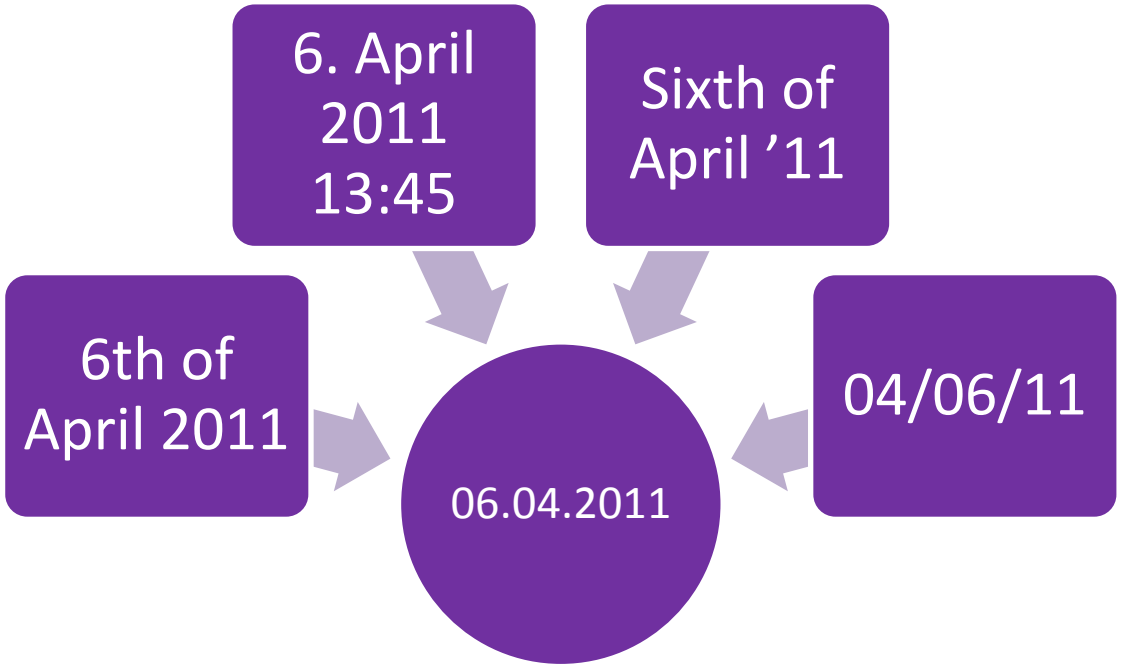
Making our content machine readable, interpretable and usable outside the context of our own institutions

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Normalization of Dates

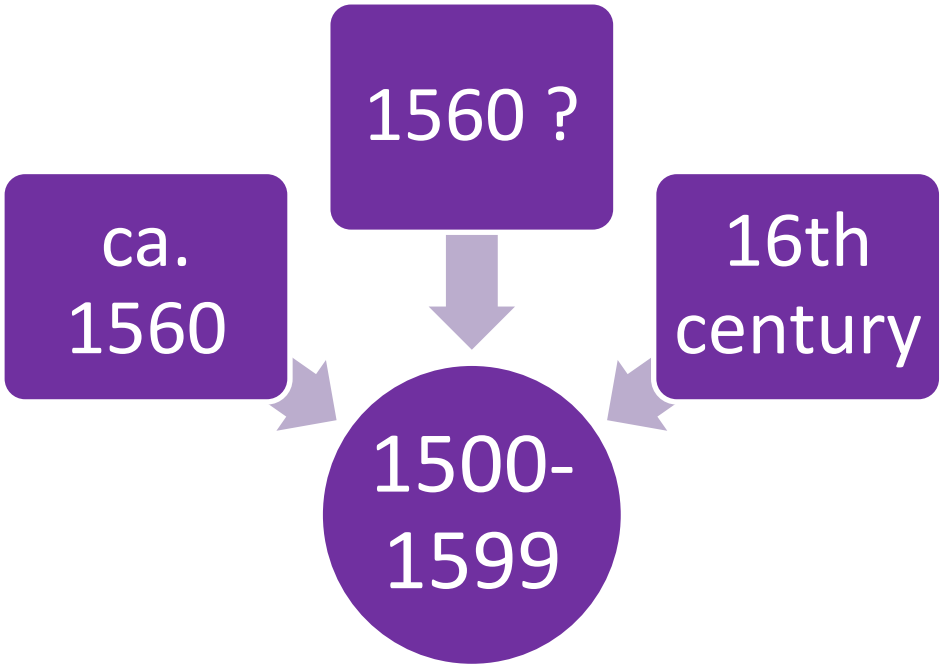


May involve generalization in the process

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Precision in Normalization?

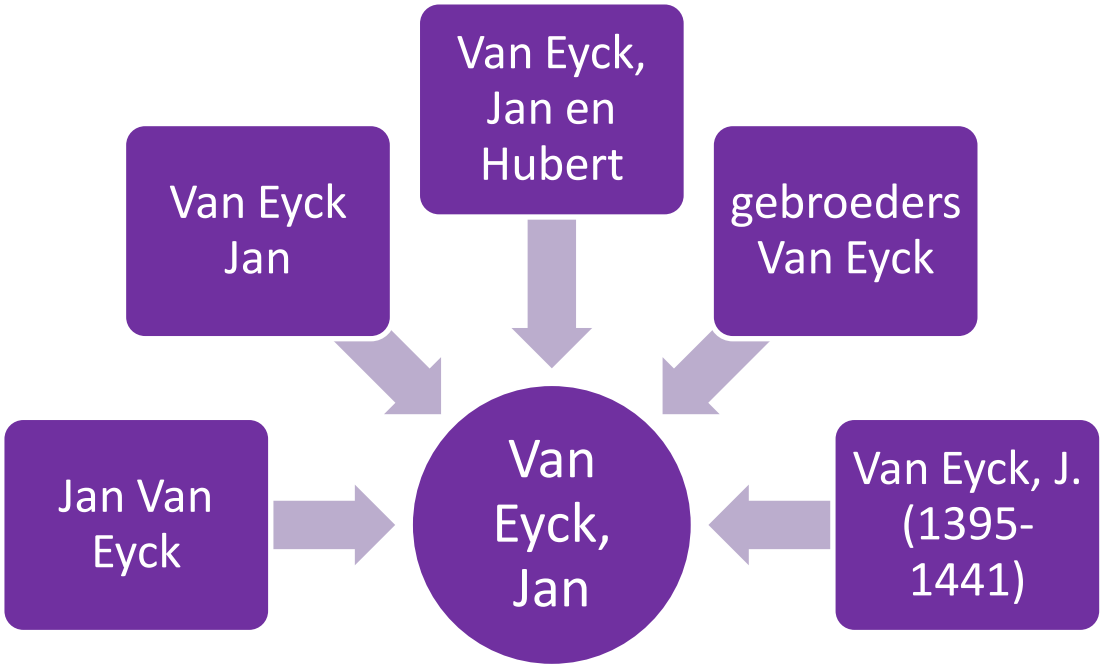


(Example from "Erfgoedplus.be", courtesy of Jef Malliet)

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Normalization of Named Entities



(Example from "Erfgoedplus.be", courtesy of Jef Malliet)

europeana
carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Metadata Enrichment

Options to improve the quality of our metadata at the time of extraction

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Enrichment: what is it?

Example

- Mapping content values to common vocabulary with defined relationships between them
- Enables vast quantities of unrelated content to be automatically linked to eachother – rendering considerable added value

Example

- Automatic keyword extraction and mapping to vocabulary
- Automatic extraction of geographic names and mapping to coordinate locations

CARARE is funded by the **European Commission's ICT Policy Support Programme**

Named Entity Extraction

Original text (from Wikipedia on Mona Lisa):

Mona Lisa (also known as La Gioconda or La Joconde) is a 16th-century portrait painted in oil on a poplar panel by Leonardo di ser Piero da Vinci during the Renaissance in Florence, Italy. The work is currently owned by the Government of France and is on display at the Musée du Louvre in Paris under the title Portrait of Lisa Gherardini, wife of Francesco del Giocondo.

After chunking/filtering

Mona Lisa / La Gioconda / La Joconde / 16th-century / Leonardo di ser Piero da Vinci / Renaissance Florence, Italy / Government of *France* / Musée du Louvre / Paris / Portrait of Lisa Gherardini / Francesco del Giocondo

Each word/chunk may be compared to controlled vocabularies such as the AAT for artist names or Geonames for place names in order to extract machine-understandable meaning from a text

Enrichment Issues

Consider the following text:

"...it was almost as bad as in the 1920s..."

"...in the Iron Age..."

Are the 1920s really relevant to this text?

When was the Iron Age?

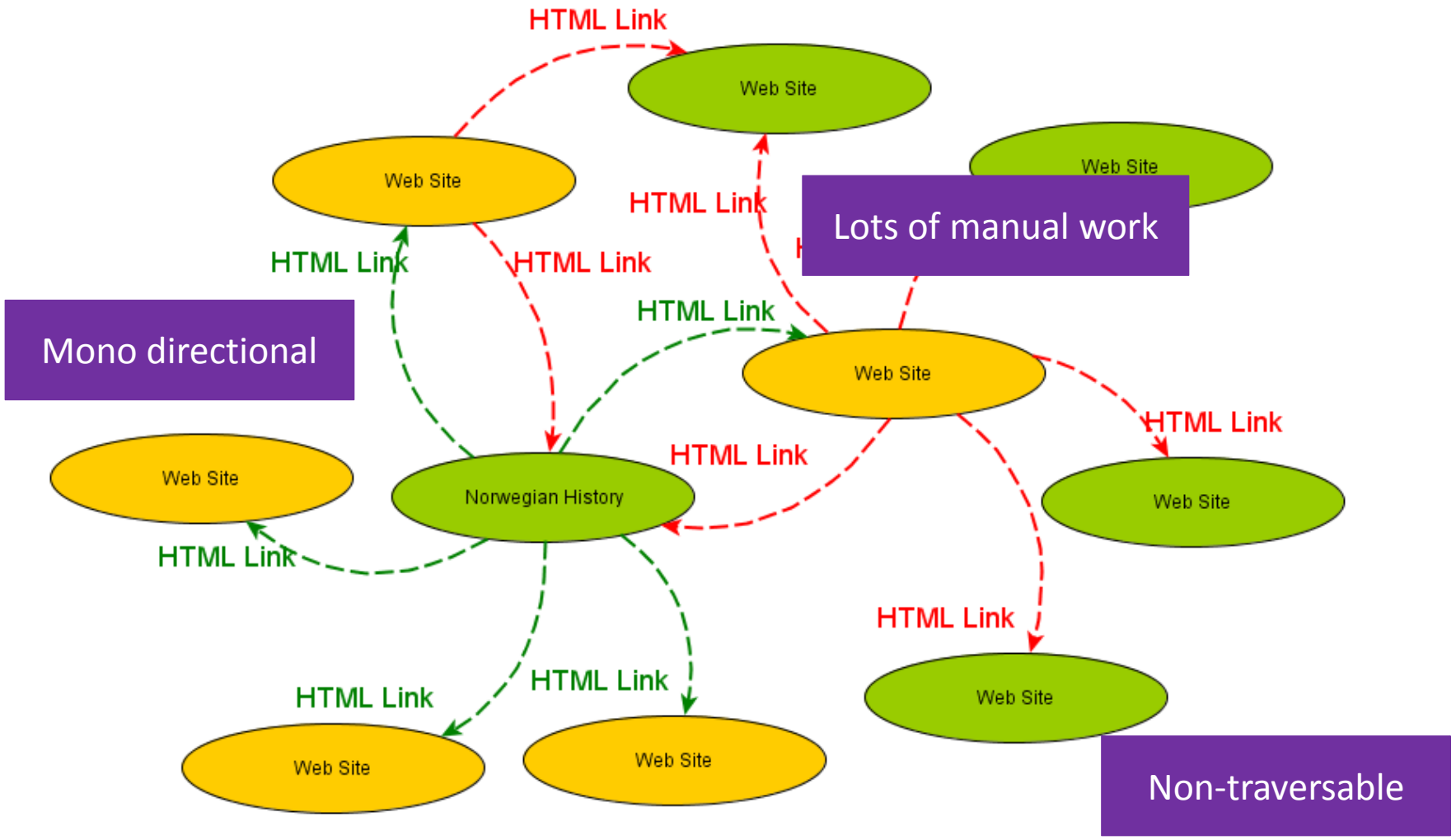
Consider the following text:

"..they were making a plot..."

Are 'they' planning to overthrow the government – or are 'they' working on establishing a cadastral parcel of land?

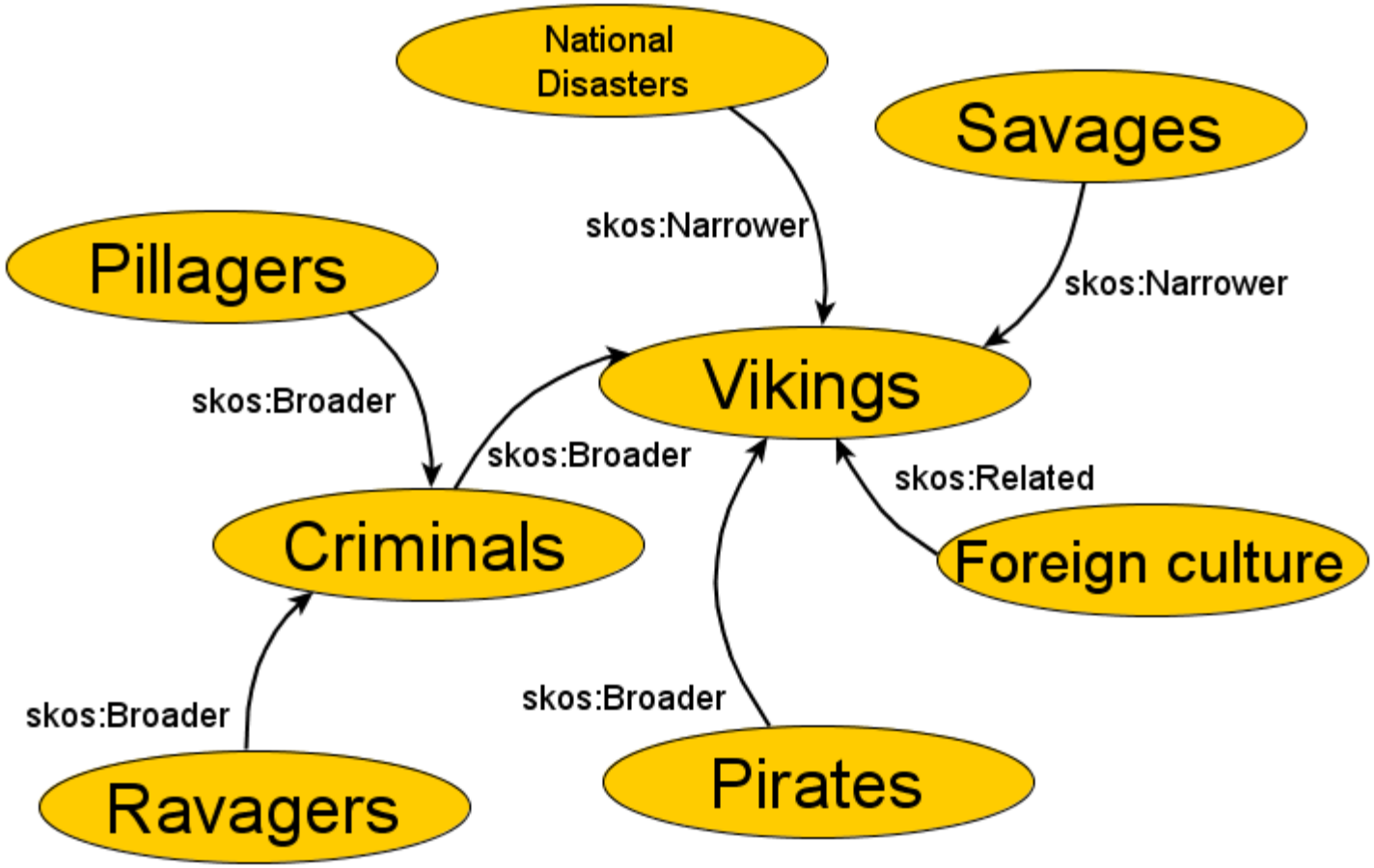
Without manual interpretation or additional contextual knowledge, it is in many cases difficult to determine the exact meaning of an extracted phrase/keyword

Manually established relationships



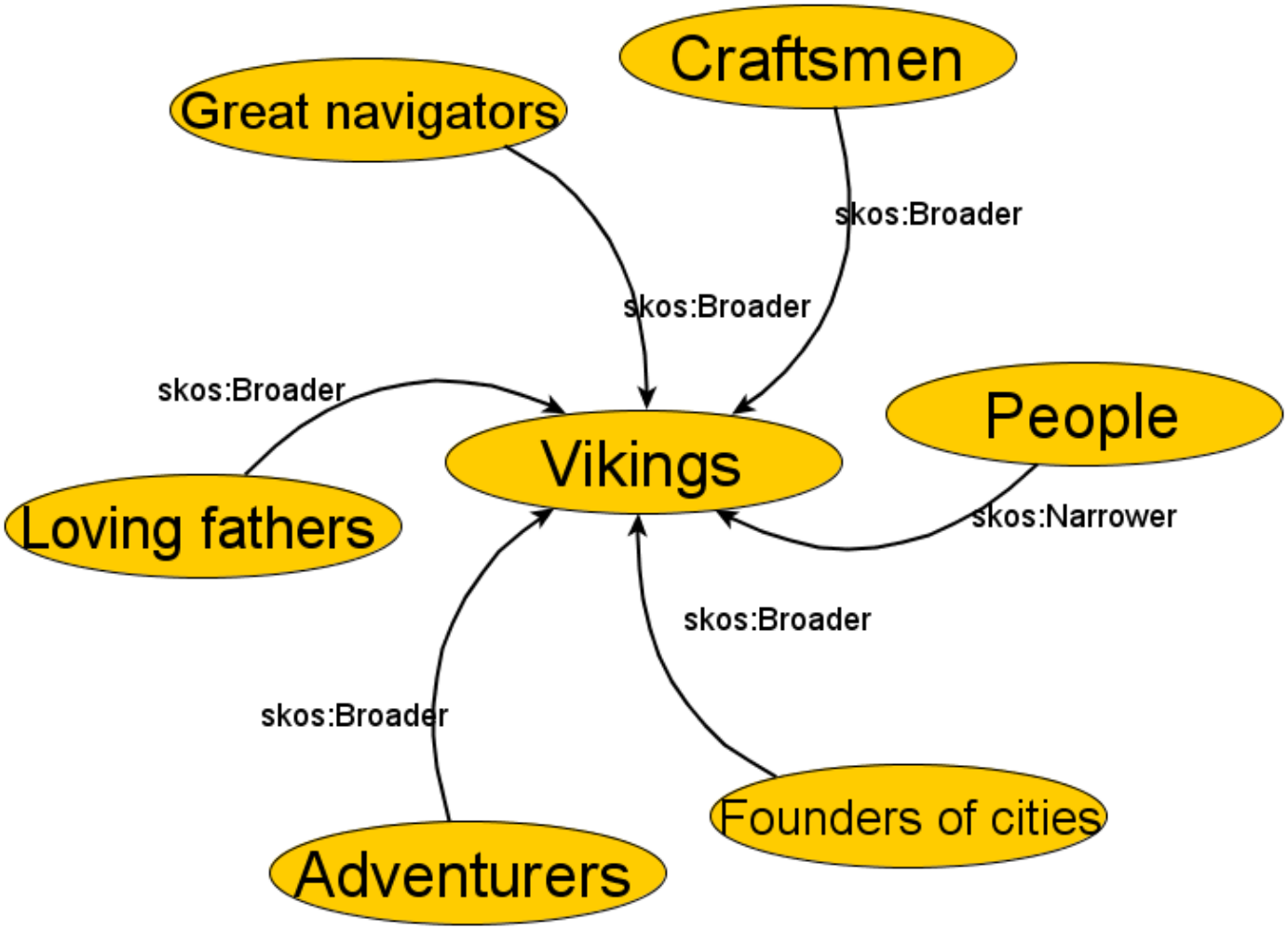
CARARE is funded by the **European Commission's ICT Policy Support Programme**

Irish vocabulary on Vikings



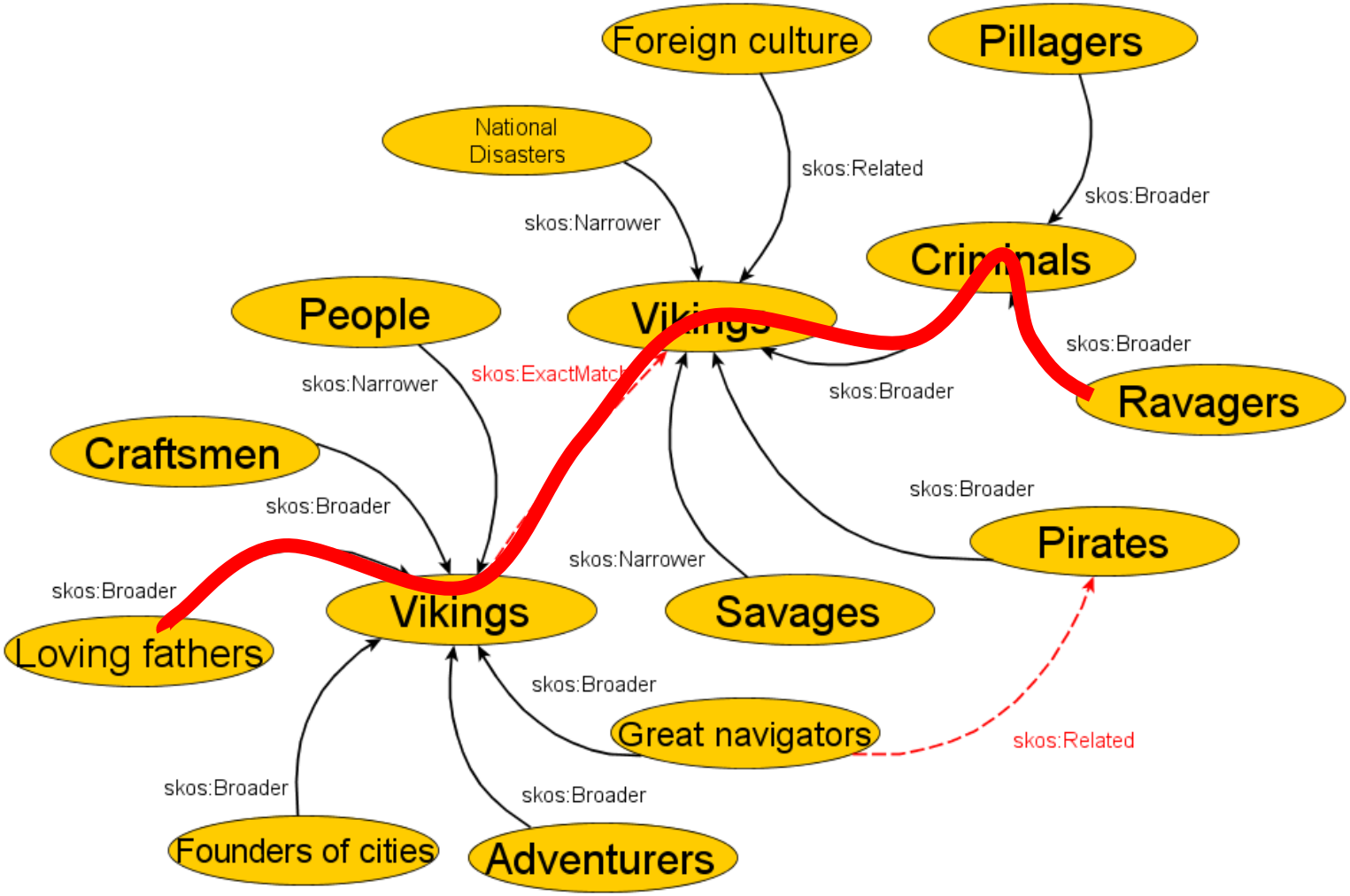
CARARE is funded by the **European Commission's ICT Policy Support Programme**

Norwegian vocabulary on Vikings



CARARE is funded by the **European Commission's ICT Policy Support Programme**

Mapped vocabularies – Semantic Graphs



CARARE is funded by the **European Commission's ICT Policy Support Programme**

Transformation Examples and Exercises

europeana

carare project

CARARE is funded by the **European Commission's ICT Policy Support Programme**

SQL XML Publishing Functions

xmlelement()	Creates an XML element, allowing the name to be specified.
xmlattributes()	Creates XML attributes from columns, using the name of each column as the name of the corresponding attribute.
xmlroot()	Creates the root node of an XML document.
xmlcomment()	Creates an XML comment.
xmlparse()	Parses a string as XML and returns the resulting XML structure.
xmlforest()	Creates XML elements from columns, using the name of each column as the name of the corresponding element.
xmlconcat()	Combines a list of individual XML values to create a single value containing an XML forest.
xmlagg()	Combines a collection of rows, each containing a single XML value, to create a single value containing an XML forest.

CARARE is funded by the **European Commission's ICT Policy Support Programme**

SQL XML example

```
select xmlelement(name "CustomerProj",
   xmlforest(c.CustId, c.Name as CustName, p.ProjId, p.Name as ProjName))
from Customers c, Projects p
where p.CustId=c.CustId
order by c.CustId
```

```
select
  xmlelement(name customer,
    xmlattributes(c.CustId as id),
    xmlforest(c.Name as name, c.City as city),
    xmlelement(name projects,
    (select xmlagg(xmlelement(name project,
       xmlattributes(p.ProjId as id),
       xmlforest(p.Name as name)))
      from Projects p
     where p.CustId=c.CustId))) as "customer-projects"
from Customers c
```

Source: http://www.stylusstudio.com/sqlxml_tutorial.html

Named Entity Extraction

Consider the following text:

"…CARARE brings together heritage agencies and organisations, archaeological museums and research institutions and specialist digital archives from all over Europe to establish a service that will make digital content for Europe's unique archaeological monuments and historic sites interoperable with Europeana. It aims to add the 3D and Virtual Reality content to Europeana…"

The following matches are found:

Two occurences of the word 'Europe' assumed to be a geograhical location

One occurrence of the word 'Virtual Reality' assumed to be an organization (?)

Try an online version of a geographical names parser:
http://geoparser.digmap.eu/