Machine-processable vocabularies: First steps towards a Semantic Web

Thomas Baker, Fraunhofer-Gesellschaft

INETBIB 2002, Goettingen

19 September 2002

Semantic Web vision

- Vision (Tim Berners-Lee) and W3C Activity
- Motivation: integrated access to a diversity of resources on the Web
- Simple linked data model
  - Webs of information about related things using explicit statements following a common model
- XML as a universal file format
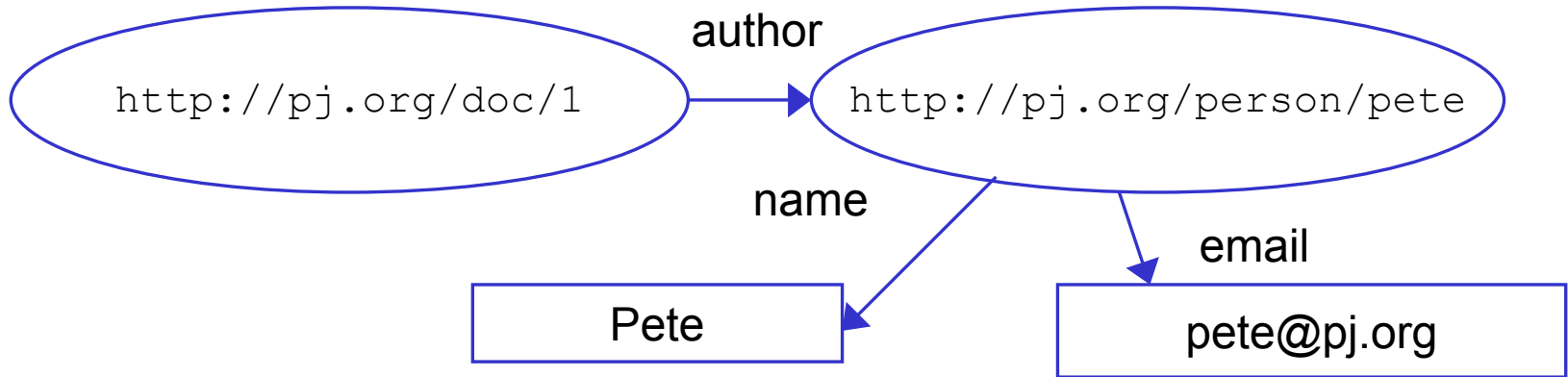- URIs as unique identifiers

Semantic Web vision
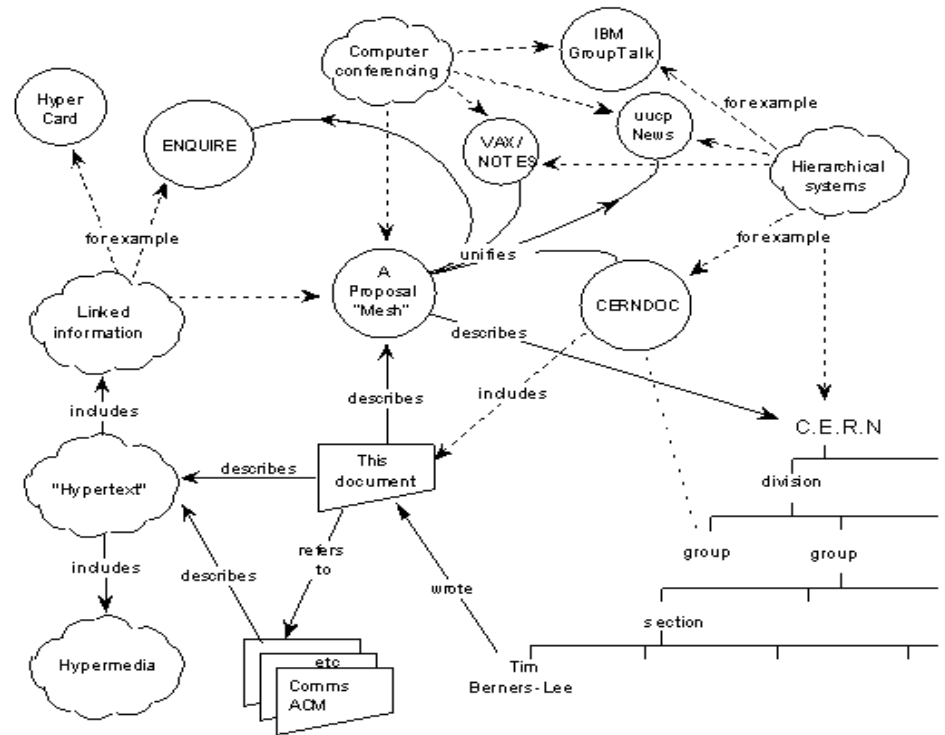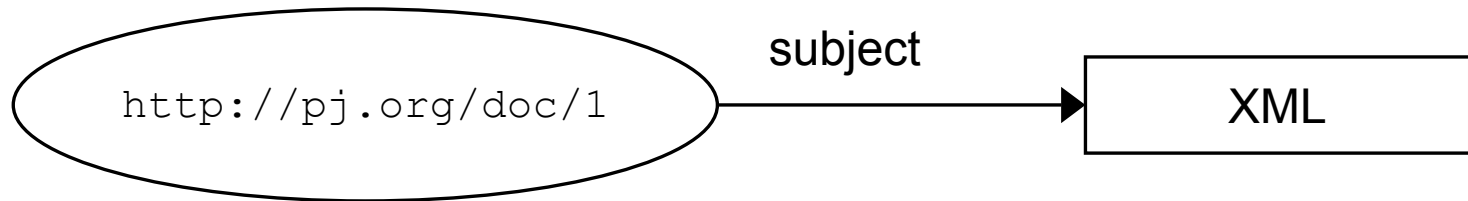
URIs as anchors for merging data

- Uniform Resource Identifiers (URIs)
  - Identify resources (people, information), e.g. http://purl.org/net/sbp/
  - Identify metadata terms used to describe those resources, e.g. http://purl.org/dc/terms/dateSubmitted
- URIs as fixed points for merging data
- Hoped-for market for aggregation, data merging, annotation, and filtering services

First source


```
<rdf:RDF xmlns:uc="http://www.ukoln.ac.uk/core/">
  <rdf:Description about="http://pj.org/doc/1">
    <uc:author>
      <rdf:Description about="http://pj.org/person/pete">
        <uc:name>Pete</uc:name>
        <uc:email>pete@pj.org</uc:email>
      </rdf:Description>
    </uc:author>
  </rdf:Description>
</rdf:RDF>
```
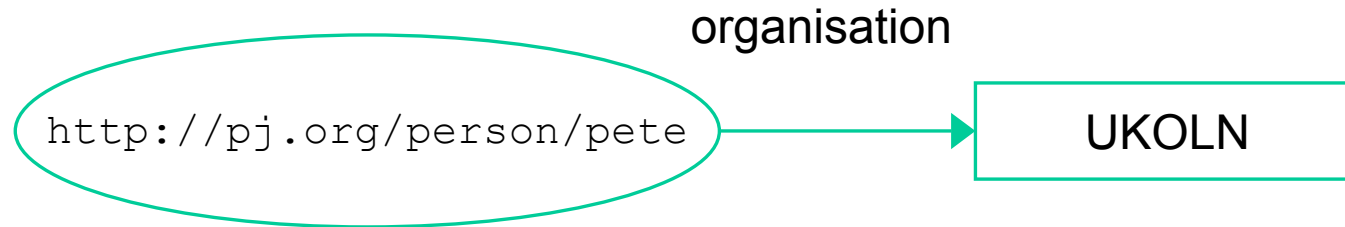
Second source


```
<rdf:RDF xmlns:uc="http://www.ukoln.ac.uk/core/">
  <rdf:Description about="http://pj.org/doc/1">
    <uc:subject>XML</uc:author>
  </rdf:Description>
</rdf:RDF>
```
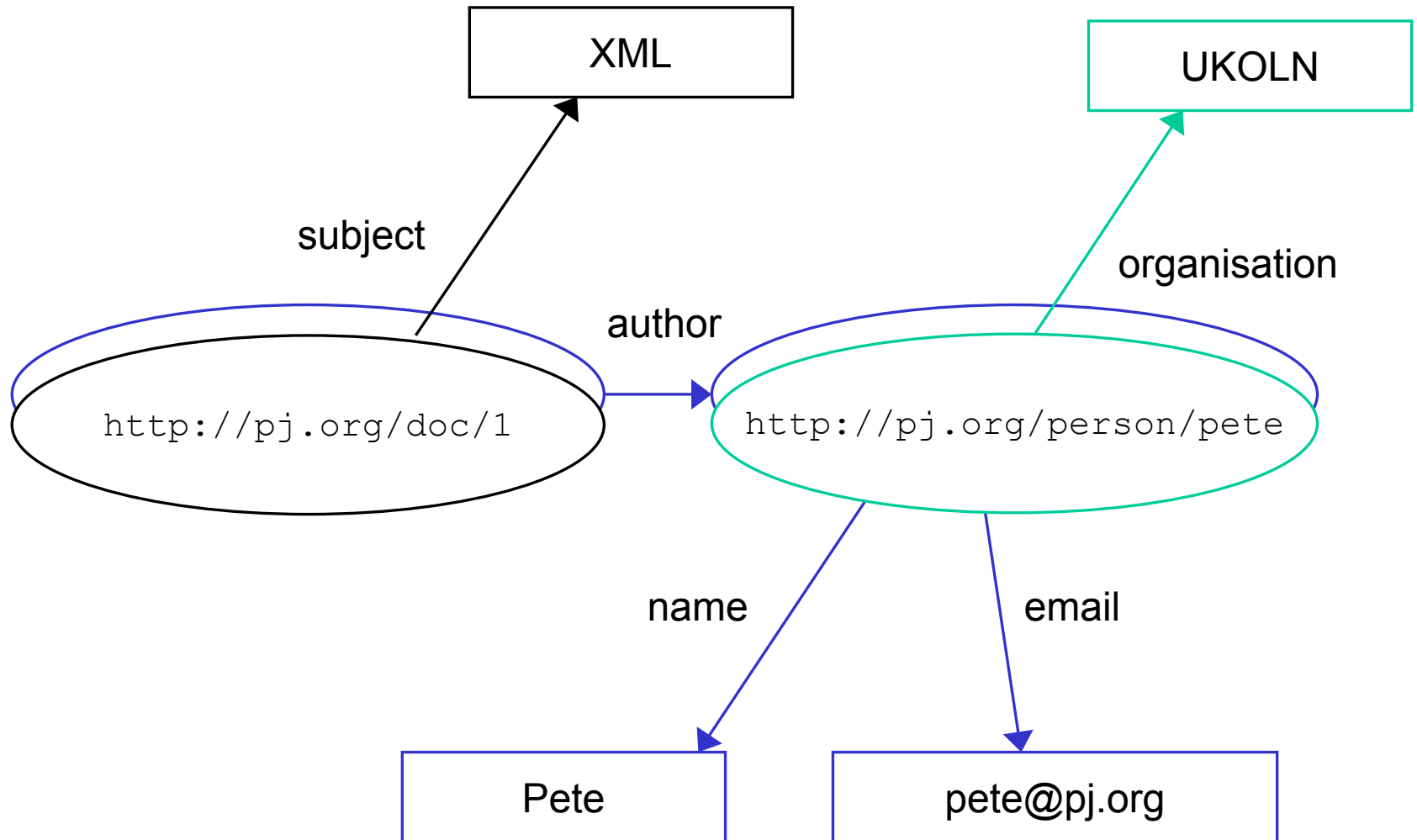
Third source

organisation


```
<rdf:RDF xmlns:uc="http://www.ukoln.ac.uk/core/">
  <rdf:Description about="http://pj.org/person/pete">
    <uc:organisation>UKOLN</uc:organisation>
  </rdf:Description>
</rdf:RDF>
```

Three descriptions merged

```
<rdf:RDF xmlns:uc="http://www.ukoln.ac.uk/core/">
  <rdf:Description
    about="http://pj.org/doc/1">
    <uc:author>
      <rdf:Description
        about="http://pj.org/person/pete">
        <uc:name>Pete</uc:name>
        <uc:email>pete@pj.org</uc:email>
        <uc:organisation>UKOLN</uc:organisation>
      </rdf:Description
    </uc:author>
    <uc:subject>XML</uc:subject>
  </rdf:Description>
</rdf:RDF>
```
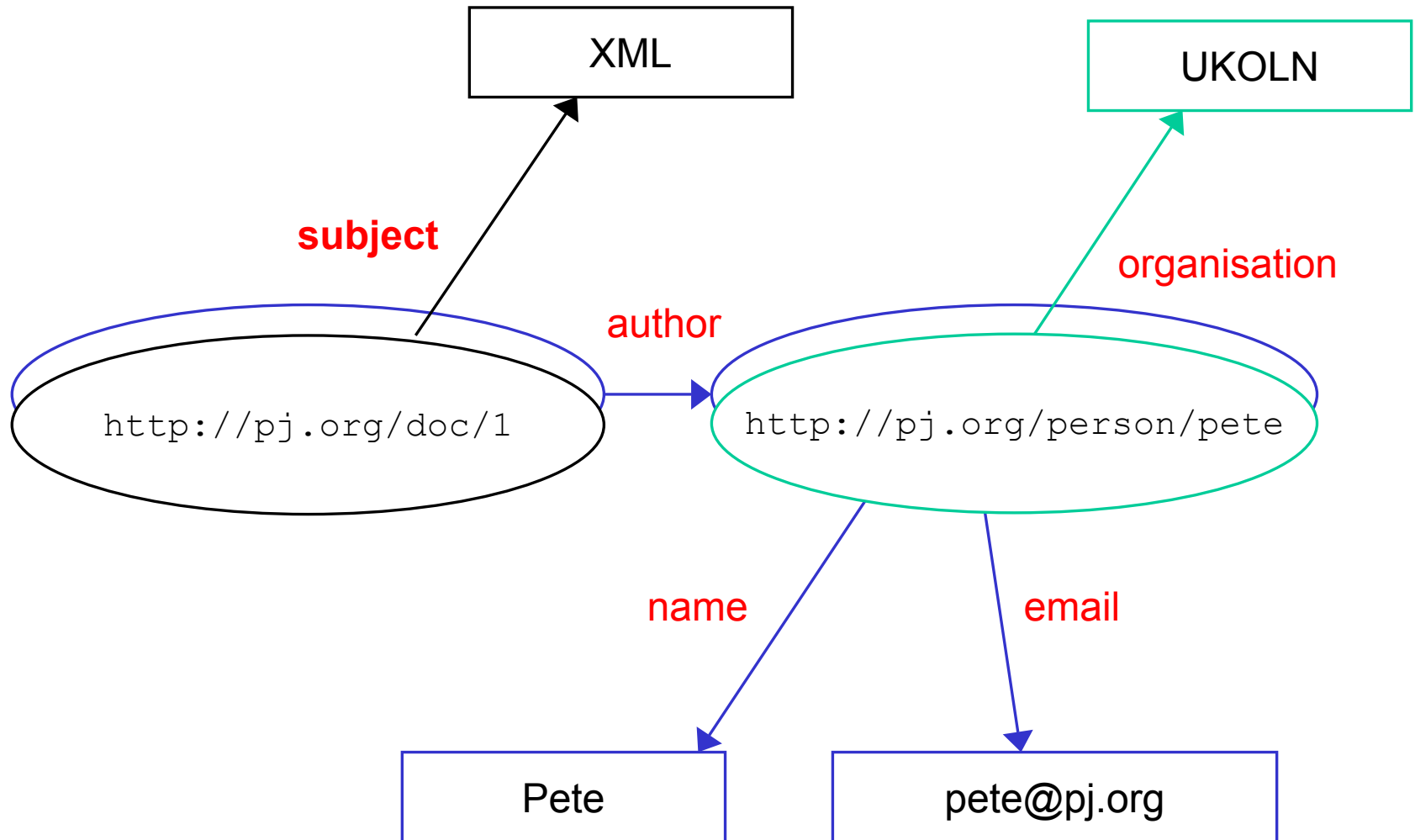
Thanks to Pete Johnston, Rachel Heery, and Andy Powell of UKOLN

Three descriptions merged

The descriptive terms themselves…

…can be URIs…

XML

http://purl.org/dc/elements/
1.1/subject

http://pj.org/doc/1

…to enable other types of data merging

XML

http://purl.org/dc/elements/1.1/subject

http://purl.org/dc/elements/1.1/subject

http://pj.org/doc/1

http://qk.org/doc/2
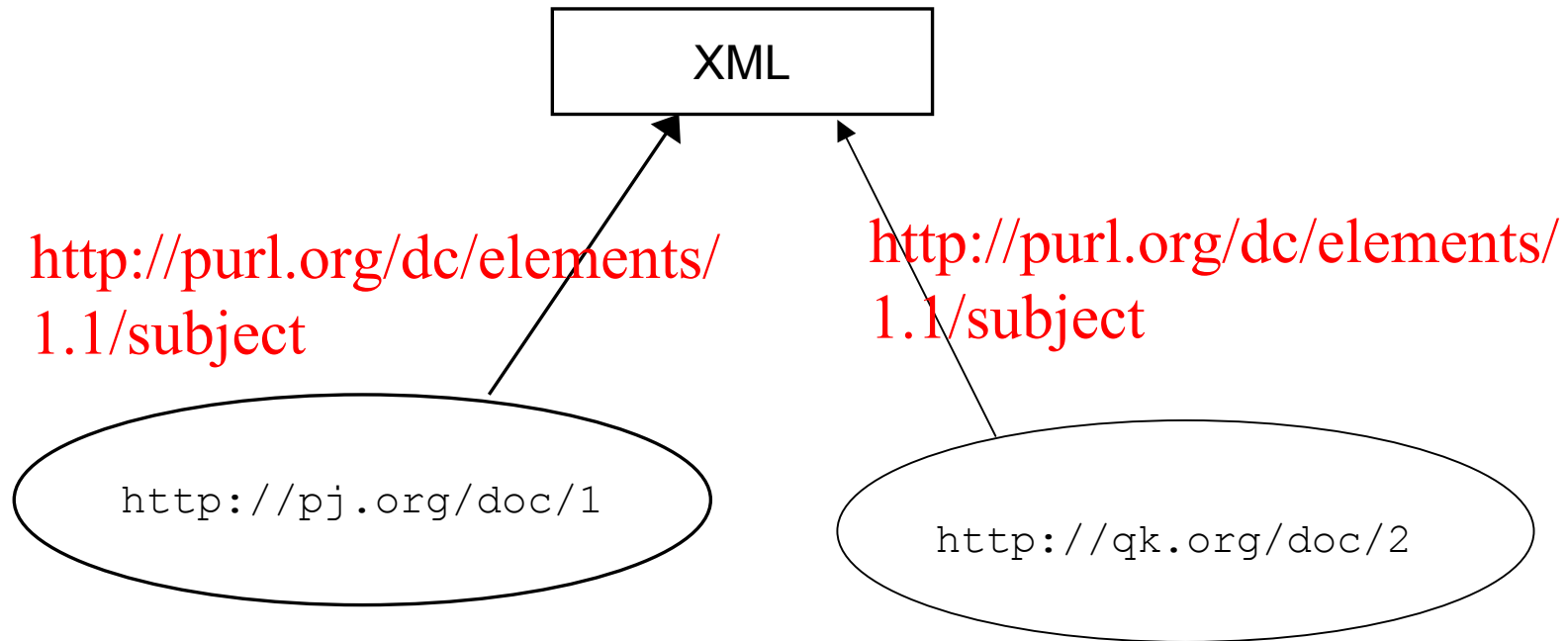
Here: two documents at two different organizations both have the **subject** "XML".

Goal: *Partial* understanding

- To share data between programs and resources designed independently
  - Essential trait of a massively distributed Web
  - Incorporate and re-purpose data for unanticipated uses
  - Communication among diverse communities on basis of *partial, imperfect* understanding
- Assumption: tolerate inconsistency and errors!
  - Ignore the ones you don't understand
  - On the Web, "Error 404: File not found", but unchecked exponential growth

To do this, "elements" need URIs

- "Elements": metadata terms used to describe things
  - People: name, email address, shoe size…
  - Books: title, author, ISBN number, publisher…
  - Computer files: format, size…
- Declared and maintained by a diversity of organizations

Dublin Core Namespace Policy

- All DCMI metadata terms are given unique identity within three namespaces:
    - http://purl.org/dc/elements/1.1/ - the legacy DC-15
    - http://purl.org/dc/terms/ - all other elements/qualifiers
    - http://purl.org/dc/dcmitype/ - a Type vocabulary
    - Example: http://purl.org/dc/elements/1.1/title
- Policy on long-term stability of namespace URIs
    - Changes not substantially "semantic" (i.e., corrections) will not result in change of namespace URIs
    - "Semantic" changes must trigger a change of name
    - Version turnover of a "document management" nature will have no effect on namespace URIs

DCMI Usage Board

- DCMI term set must evolve as implementors coin new terms and usage patterns emerge
- Usage Board reviews proposals for new metadata terms (elements, refinements, encoding schemes, Type terms)
  - Evaluates proposals in light of grammatical principle, usefulness, clarity of definition, overlap with existing terms
- Reviews application profiles that are based substantially (though not exclusively) on Dublin Core
- Tiered model of approval status: conforming, recommended, obsolete

Shared conventions for declaring elements?

- CORES Interoperability Forum (EU)
  - Interoperability survey including MPEG, DOI, MARC21, GILS, Dublin Core, CERIF, OASIS, W3C, ONIX, IEEE-LOM
  - Summit meeting in November 2002
  - Do we have elements? Are they comparable? Should we identify them in a common way? Do we share an understanding of what that would mean?
  - http://www.cores-eu.net/

Registries: value-added services needing to cite elements

- Web-based dictionaries of metadata terms
  - Within standards-using communities (eg Dublin Core)
  - Within communities of practice (eg, FAO for development information; MEG for education)
  - Within large organizations (eg, Microsoft intranet)
  - For crosswalks (MetaForm in Goettingen)
  - To enable metadata conversion services (eg, CNRI Distributed Metadata Service)

DCMI Registry

- Prototype database based for serving up term definitions, translations, and application profiles related to Dublin Core

- Plain-vanilla, good-practice application of open Web standards (e.g., RDF schemas)

- Liaison with related initiatives on defining shared conventions for declaring vocabularies on the Web (e.g., EU CORES Project)

- Planning for machine interface to future registry-using applications

Thomas.Baker@bi.fhg.de