

Automatically generating citation text from queries

Gianmaria Silvello

Department of Information Engineering,
University of Padua, Italy

gianmaria.silvello@unipd.it

<http://www.dei.unipd.it/~silvello/>

Web Conference Data Citation WG
Webinar, 07 November 2017

Outline

What's data citation: a very brief introduction

Citing RDF: The eagle-i case

Citing XML: The digital archives case

Citing RDB: The IUPHAR/BPS case

Conclusions

What's Data Citation

Check [\[JASIST2017b\]](#) for a survey about theory and practice of data citation

Publication is changing

- Information is increasingly published on the web.
- Much of this information is in curated databases – crowd- or expert-sourced data
- These datasets are complex, structured, and evolving, and contributors need to be acknowledged



Data citation desiderata

- The generation of human- and machine-readable citations should be automatic
- Cited data should be uniquely identified: e.g., DOI
- Citing data should be easy: click, generate, copy and paste
- Setting up and maintaining a citation system should require low (no) effort to data creators/curators

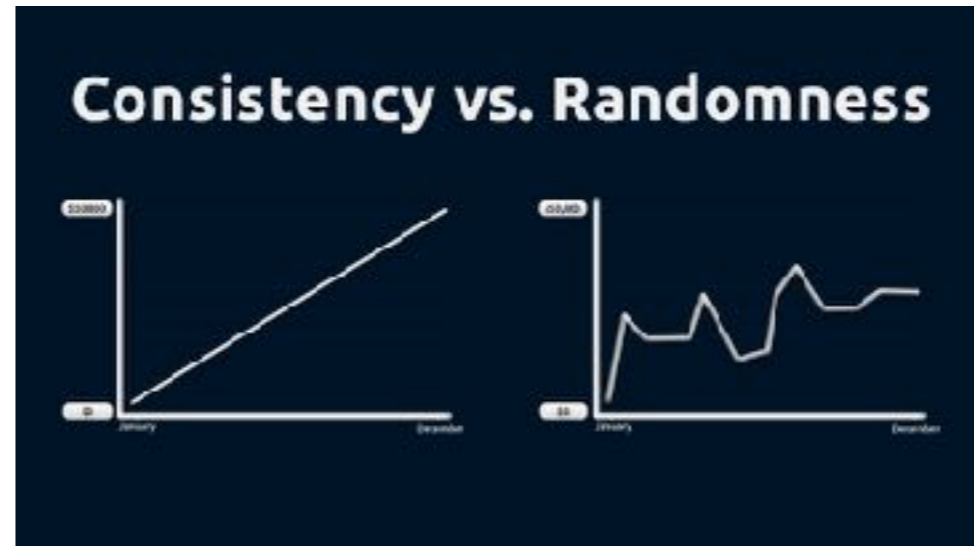
Citation snippets

We've focused on the automatic creation of citation texts/snippets

- A collection of information: authors, title, date, etc. and some kind of access mechanism
- Not exactly provenance
- Self contained, immutable (to within some choice of format)
- Needed for a variety of reasons: kudos, currency, authority, recognition, access...

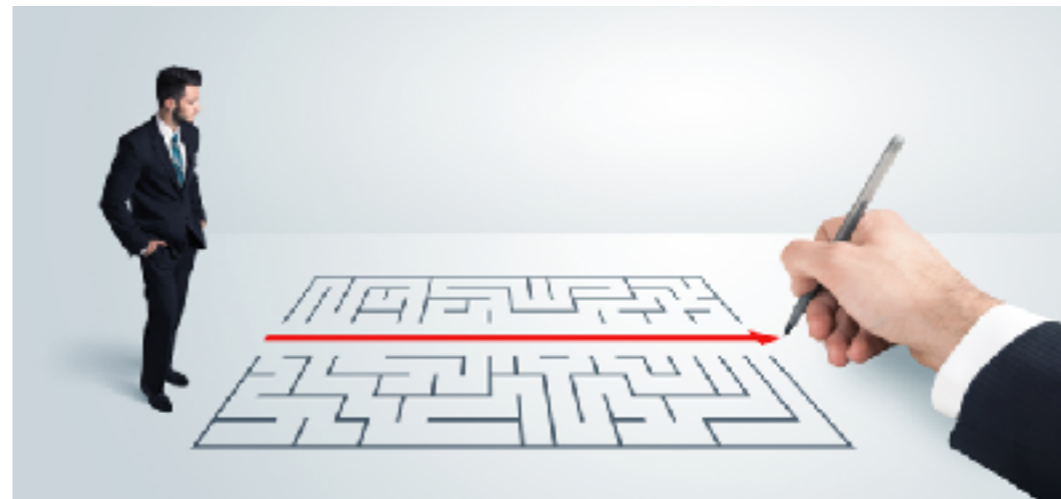
Why citation snippet generation should be automatic?

- Consistency



<https://www.investorsunderground.com/wp-content/uploads/2015/05/consistencyfeatured.jpg>

- Simplicity



<https://speckyboy.com/wp-content/uploads/2016/01/simplicity-equals-sanity.png>

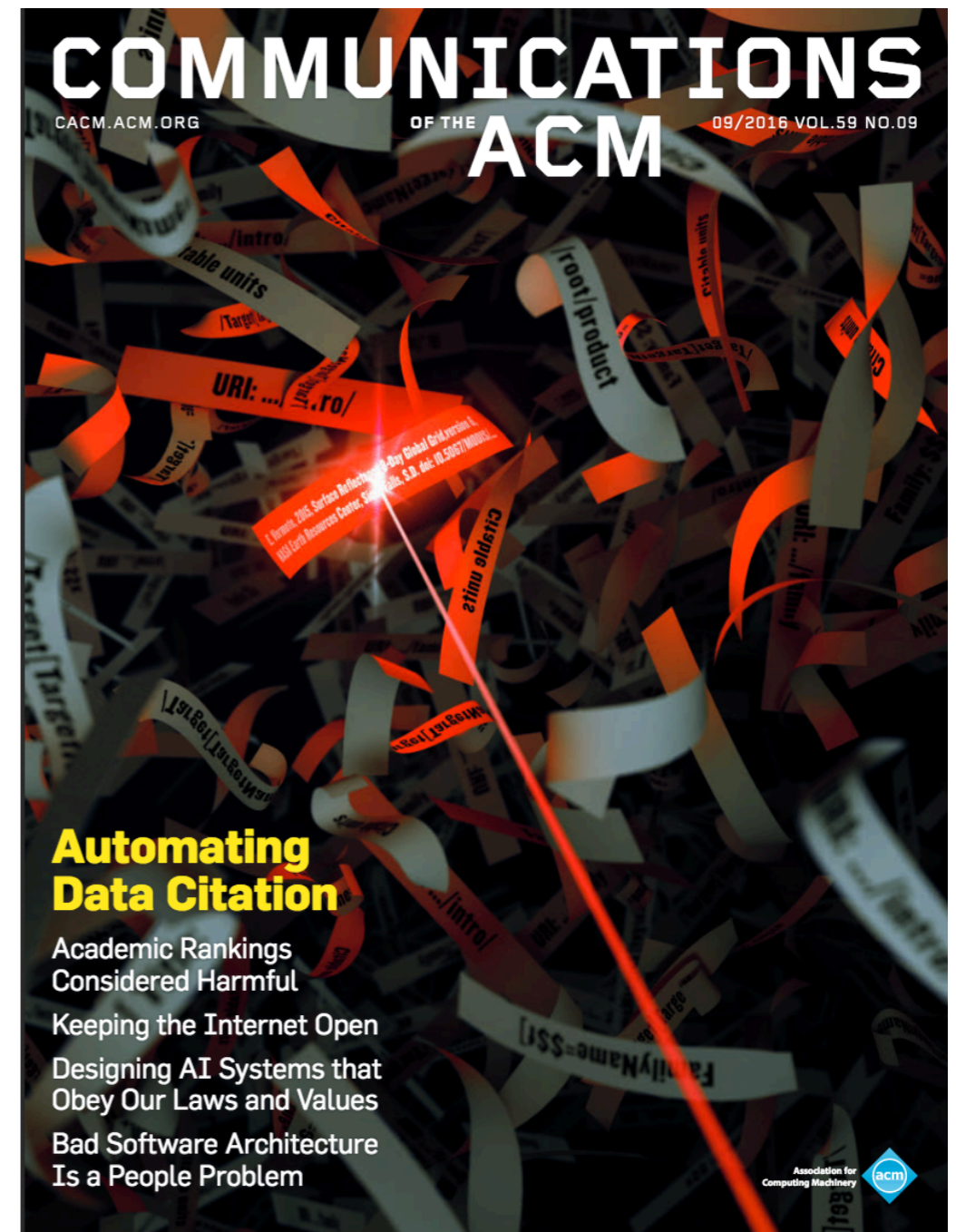
- Completeness



https://newsletter.echa.europa.eu/image/image_gallery?uid=b0ed090d-03ab-46b6-bdbf-39ee615dd6fe&groupId=6362380&t=1442480673121

Data citation as a computational problem

- Principles and standards for data citation are unlikely to be used unless the process of extracting information is coupled with that of providing a citation for it.
- We need to automatically generate citations as the data is extracted.
- Data citation is a computational problem.



Buneman, Davidson, Frew:
Why data citation is a computational problem.
[Commun. ACM 59\(9\): 50-57 \(2016\)](#)

Data models heterogeneity

Models

Relational model



eXtensible
Markup
Language **X
M
L**

Applications



Citing RDF Using Named Graphs/Views

Application: Eagle-i

Check DLib2015 and JCDL2017

Joint work with Abdu Alawini, Leshang Chen and Susan Davidson

Eagle-i

- A “resource discovery” tool built to facilitate translational science research.
- Developed by a consortium of universities under NIH funding, headed by Harvard.
- End users: researchers who wish to share information about research resources (Core Facilities, iPS cell lines, software resources).
- Data is stored and distributed as RDF files (graph database).
- Resources have a “Cite this resource” button!



Eagle-i: Cite me button



Search for resources across the eagle-i Network

Go

[Top Categories](#) | [Explore All](#)

[ABOUT](#) [GET INVOLVED](#) [NEWS + EVENTS](#) [FAQ](#) [CONTACT US](#) [HELP](#)

[< Back to Search Results >](#)

Mouse anti-mouse Gamma-protocadherin-C4

Monoclonal antibody reagent ⓘ

Send message to
resource contact

Cite this resource



Special Collections

Reagent
Additional Name anti-PCDH-gamma-C4

Location [UC Davis/NIH NeuroMab Facility](#)

Related
Technique Immunocytochemistry ⓘ

Source Organism
Type [Mus musculus](#)

▶ Antibody Target(s)

▶ Immunogenic Material

Isotype IgG1 ⓘ

Antibody
Registry ID http://antibodyregistry.org/AB_10671301
http://antibodyregistry.org/AB_2159730

Catalog Number 73-231
75-231

Clone ID N193A/13

Exchange
Facilitator [Order from the NeuroMab Facility](#)

Website(s) http://neuromab.ucdavis.edu/datasheet/N193A_13.pdf

Eagle-i: Cite me button



Search for resources across the eagle-i Network

Go

Explore All

Mouse anti-mouse Gamma-protocadherin-C4

Monoclonal antibody reagent ⓘ

Send message to
resource contact

Cite this resource

Reagent
Additional Name anti-PCDH-gamma-C4

Location [UC Davis/NIH NeuroMab Facility](#)

Related
Technique Immunocytochemistry ⓘ

Source Organism
Type [Mus musculus](#)

▶ Antibody Target(s)

▶ Immunogenic Material

Isotype IgG1 ⓘ

Antibody
Registry ID http://antibodyregistry.org/AB_10671301

http://antibodyregistry.org/AB_2159730

Eagle-i: Cite me button

The screenshot shows the Eagle-i website interface. At the top left is the Eagle-i logo and navigation links: ABOUT, GET INVOLVED, and NEWS. A search bar is located at the top right with the text 'Search for resources across the eagle-i Network' and a 'Go' button. The main content area displays the title 'Mouse anti-mouse Gamma-protocadherin-C4' and the subtitle 'Monoclonal antibody reagent' with an information icon. Below the title are two buttons: 'Send message to resource contact' and 'Cite this resource', with the latter highlighted by an orange border. A dropdown menu is open, showing the 'eagle-i ID for this resource:' as 'http://shared.eagle-i.net/i/00000153-5d03-c8ca-518c-569b80000000'. Below the ID is a link 'Click here' for citation examples and more information, also highlighted with an orange border, and a 'Close' button. The dropdown menu is positioned over a table of resource details. The table includes fields for 'Reagent' (anti-PCDH-gamma-C4), 'Additional Name', 'Location', 'Related Technique', and 'Source Organism Type'. Below the table are expandable sections for 'Antibody Target(s)' and 'Immunogenic Material'. The 'Isotype' is listed as 'IgG1' with an information icon. At the bottom, the 'Antibody Registry ID' is provided with two links: 'http://antibodyregistry.org/AB_10671301' and 'http://antibodyregistry.org/AB_2159730'.

Send message to resource contact **Cite this resource**

Reagent anti-PCDH-gamma-C4

Additional Name

Location

Related Technique

Source Organism Type

eagle-i ID for this resource:
http://shared.eagle-i.net/i/00000153-5d03-c8ca-518c-569b80000000

Click [here](#) for citation examples and more information.

Close

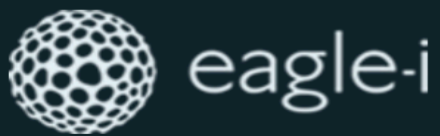
► **Antibody Target(s)**

► **Immunogenic Material**

Isotype IgG1 **i**

Antibody Registry ID http://antibodyregistry.org/AB_10671301
http://antibodyregistry.org/AB_2159730

Eagle-i: Cite me button



Search for resources across the eagle-i Network

ABOUT GET INVOLVED NEWS + EVENTS FAQ CONTACT US HELP

Citing an eagle-i Resource

Citing eagle-i resources is an easy way to give credit.

The formats suggested below provide the minimum information necessary to identify and credit the resource provider, and are designed to provide a traceable, durable, and unambiguous reference for the resource being cited. These suggestions can and should be used along with those from other resource identifiers (i.e. Antibody Registry ID, Addgene, DSHB, RRID) or from the journal publishing your work.

The screenshot shows a resource page for "APP cKO x Cre ER x APLP2 KO" in *Mus musculus*. The page includes buttons for "Request this resource" and "Cite this resource". A callout box provides the eagle-i ID and a link for citation examples. Callouts identify the resource name, eagle-i ID, eagle-i institution (Harvard University), and owning organization (Young-Pearse Laboratory).

| | |
|-------------------------------|---|
| Resource Name and Type | APP cKO x Cre ER x APLP2 KO <i>Mus musculus</i> |
| eagle-i ID | eagle-i ID for this resource: http://harvard.qa.eagle-i.net/i/0000012a-25bf-e274-f5ed-943080000002 |
| eagle-i Institution | Harvard University |
| Owning Organization | Young-Pearse Laboratory |

Organism or Virus Description: Used to study brain pathology and

Location: [Young-Pearse Laboratory](#)

Genetic alteration: [APLP2 deletion](#)
[APP cKO](#)

Close

Note that for all types, the names of Core Facilities or other ambiguously named organizations should be followed by the name of the affiliated eagle-i institution in order to disambiguate them (e.g. *Flow Cytometry Core, Montana State University* vs. *Flow Cytometry Core, Dartmouth College*).

Manual construction of the citation

The screenshot shows the eagle-i website interface. At the top, there is a search bar with the text "Search for resources across the eagle-i Network" and a "Go" button. Below the search bar is a navigation menu with links for "ABOUT", "GET INVOLVED", "NEWS + EVENTS", "FAQ", "CONTACT US", and "HELP".

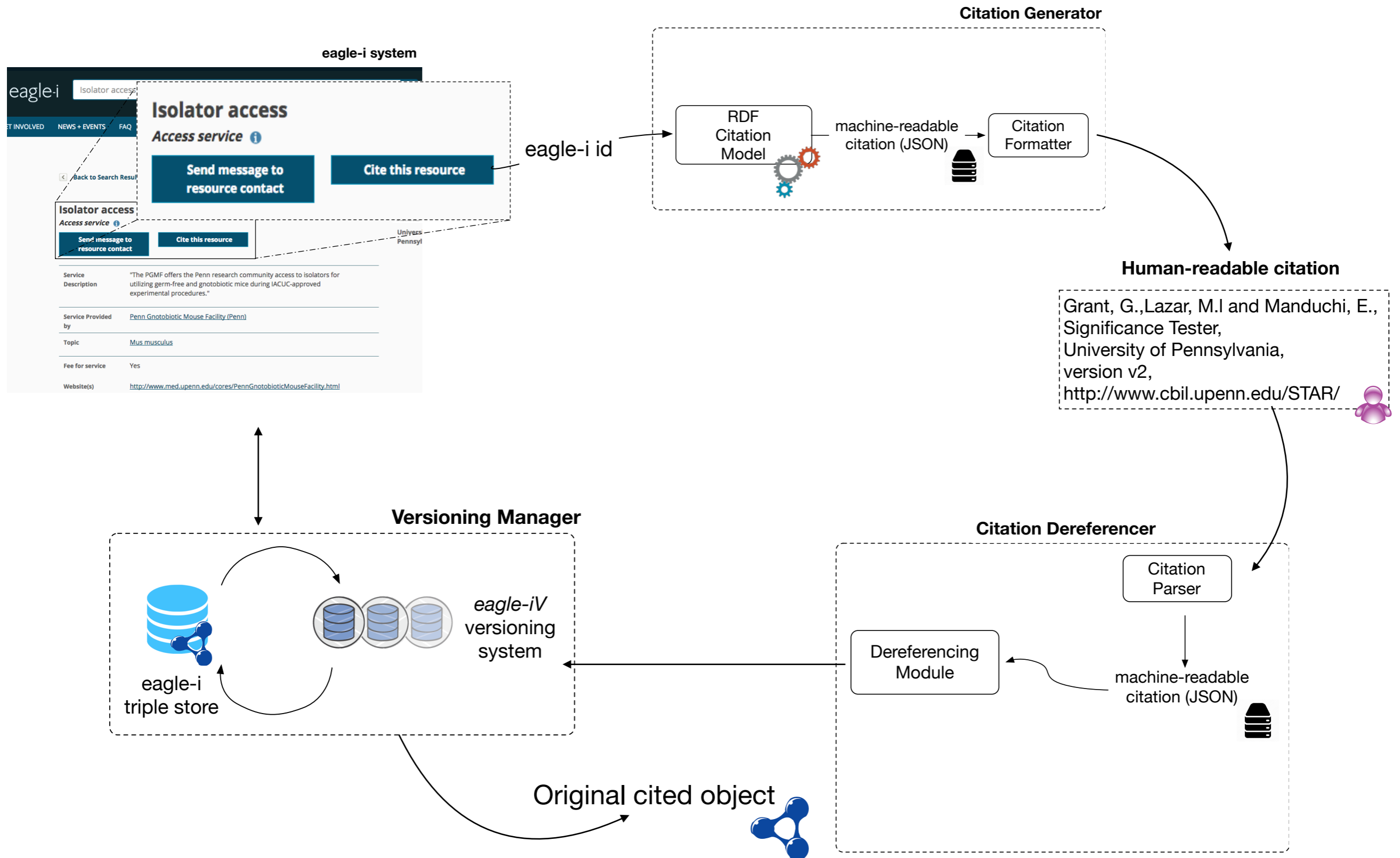
The main content area displays a resource page for "Mouse anti-mouse Gamma-protocadherin-C4". The title is highlighted with a blue box. Below the title, there are two buttons: "Send message to resource contact" and "Cite this resource". The resource details include:

- Reagent: anti-PCDH-gamma-C4
- Additional Name: (empty)
- Location: UC Davis/NIH NeuroMab Facility (highlighted with a blue box)
- Related Technique: Immunocytochemistry
- Source Organism: Mus musculus
- Type: (empty)
- Antibody Target(s): (empty)
- Immunogenic Material: (empty)
- Isotype: IgG1

On the right side, there is a "Special Collections" icon with the text "Special Collections".

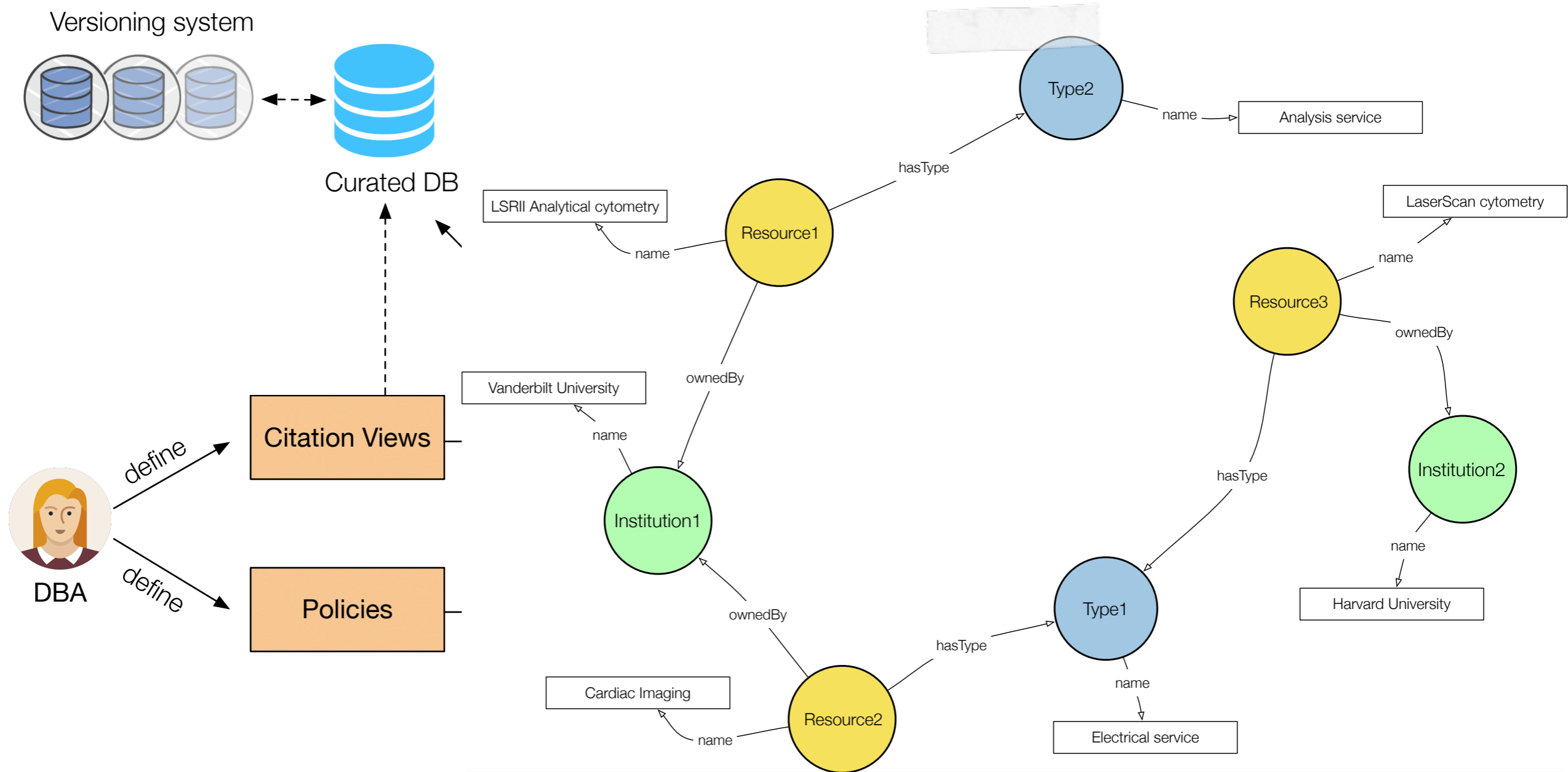
A popup window is overlaid on the bottom left, titled "Eagle-i ID". It contains the text "eagle-i ID for this resource:" followed by the URL "http://shared.eagle-i.net/i/00000153-5d03-c8ca-518c-569b80000000" (highlighted with a blue box). Below the URL, it says "Click [here](#) for citation examples and more information." and a "Close" button.

The eagle-i citation service

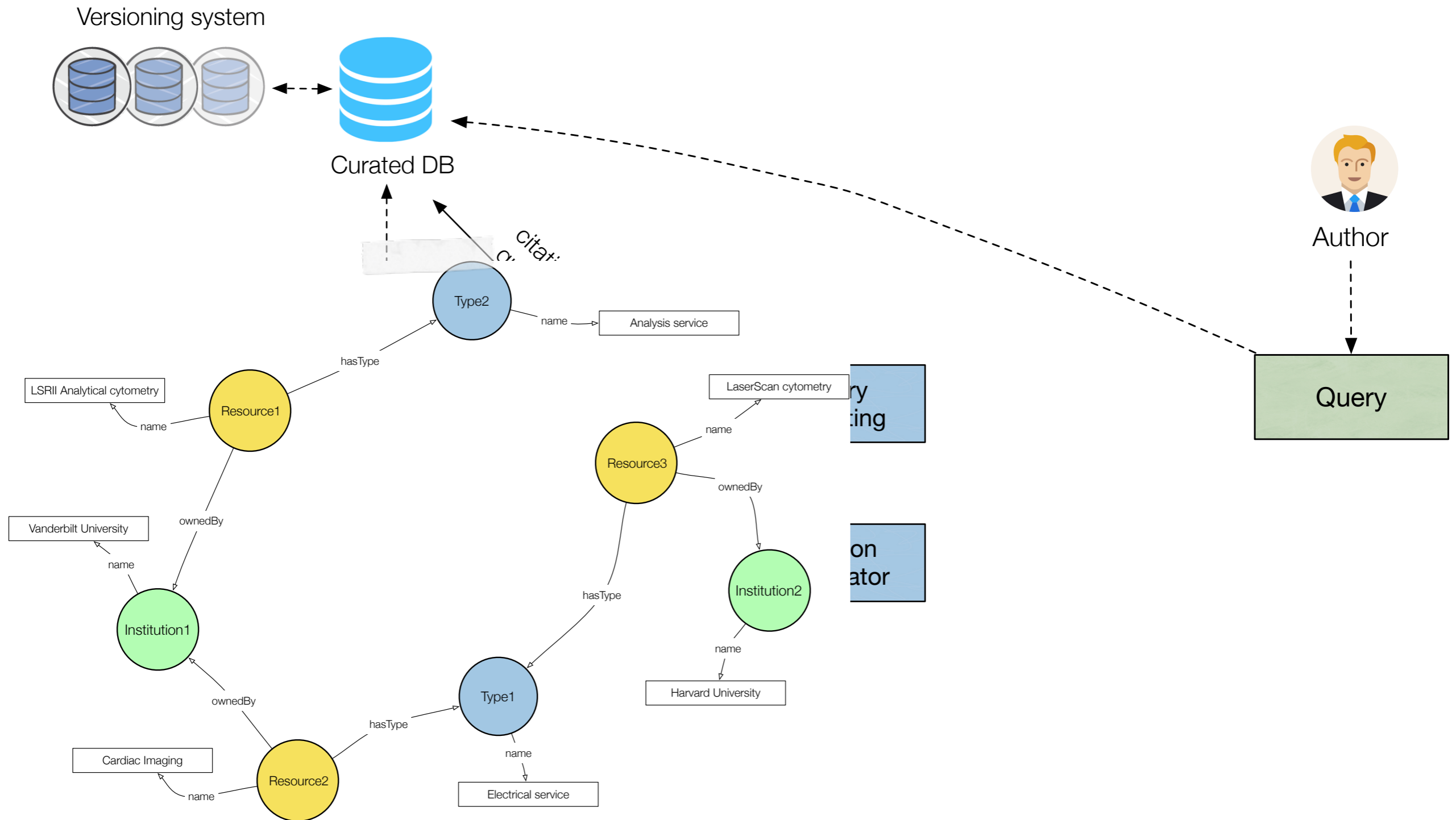


The citation system

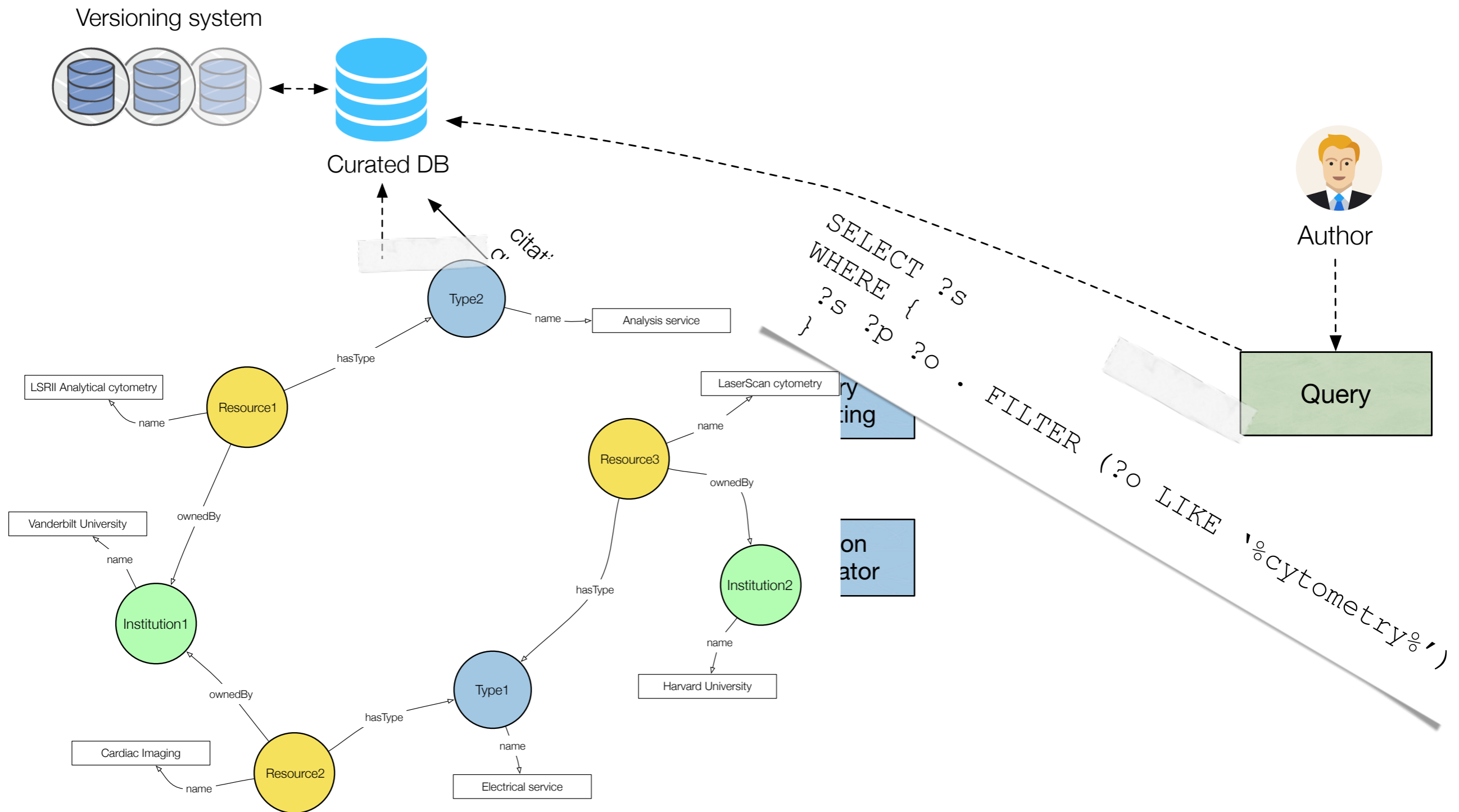
Creating a citation



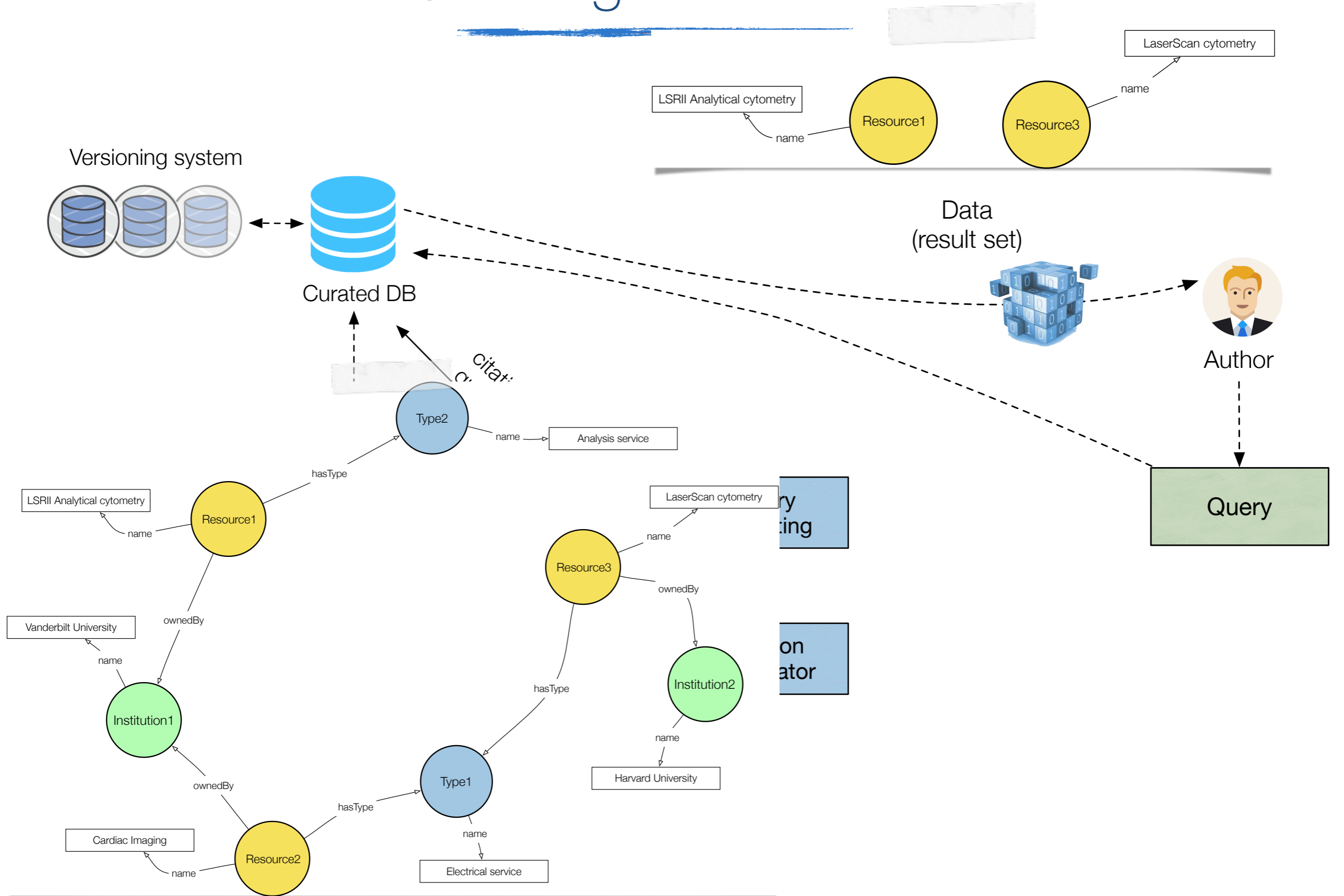
Creating a citation



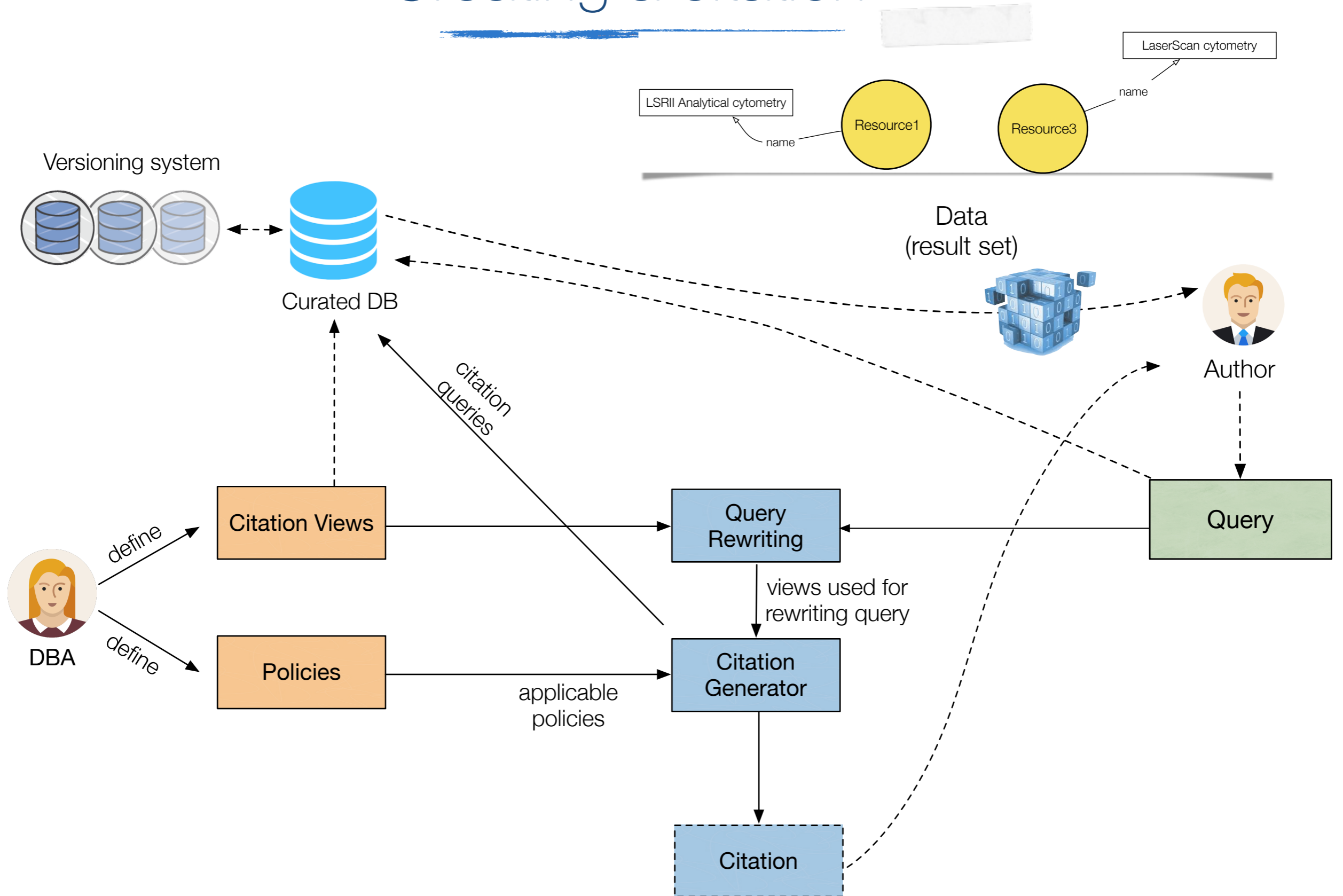
Creating a citation



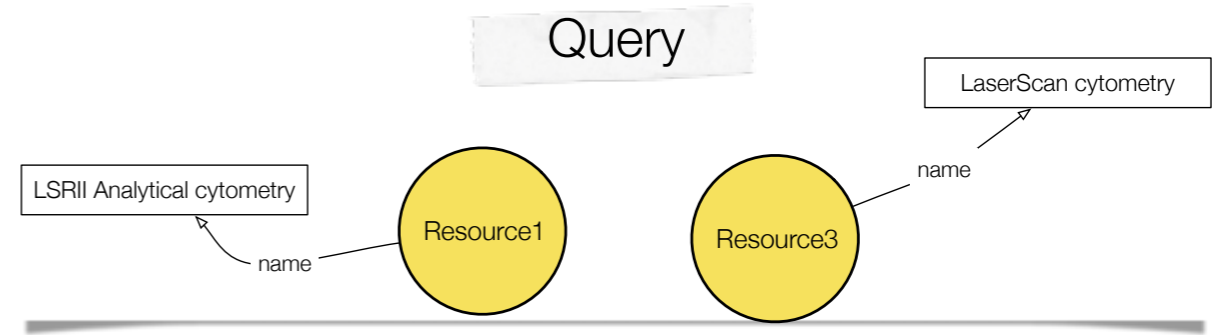
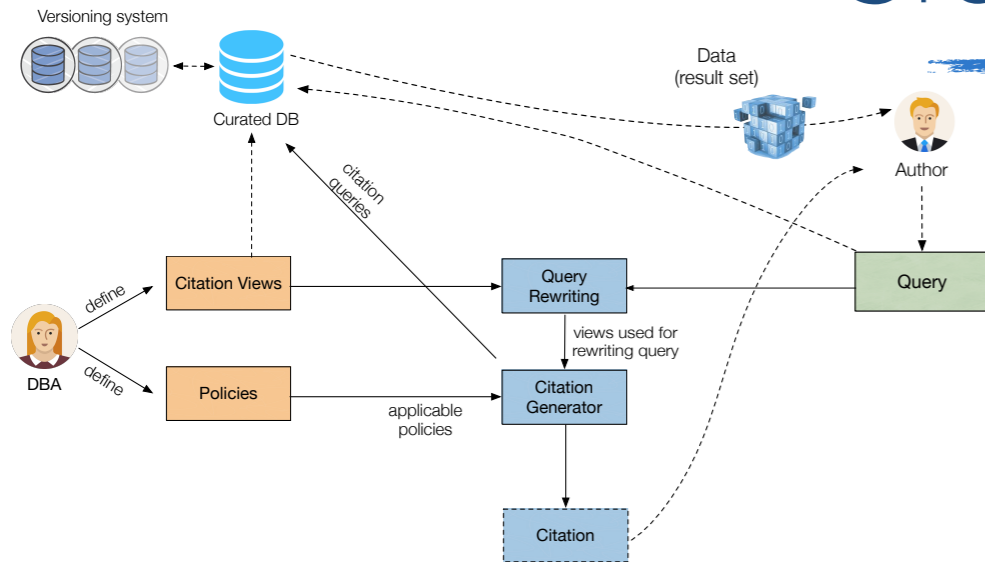
Creating a citation



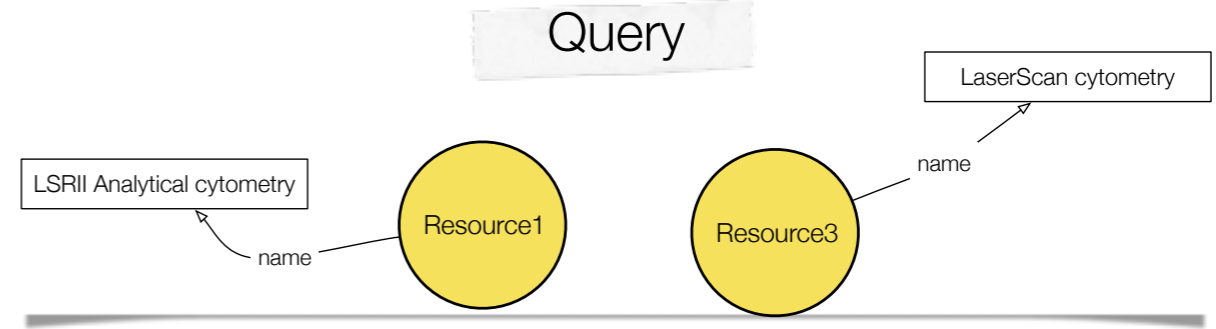
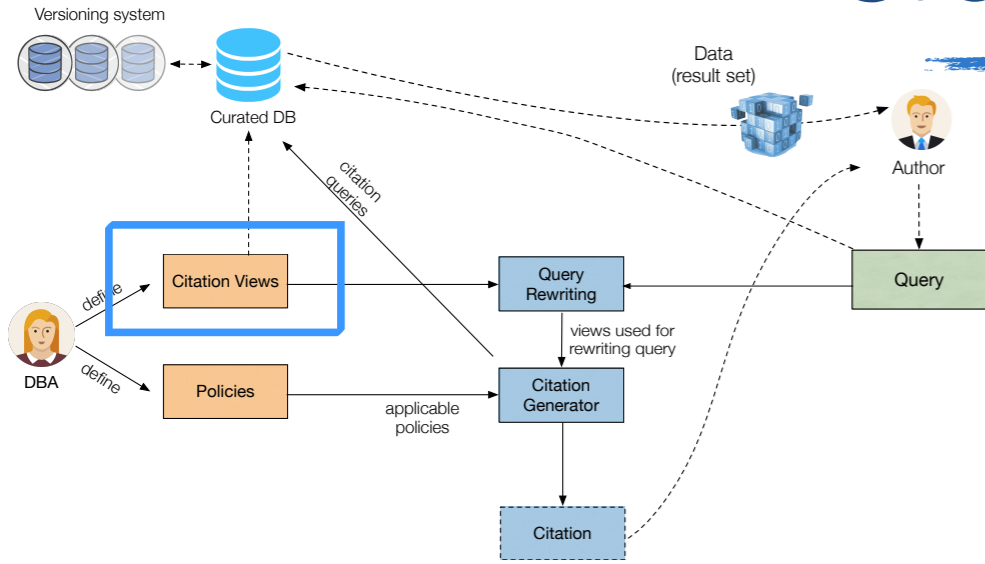
Creating a citation



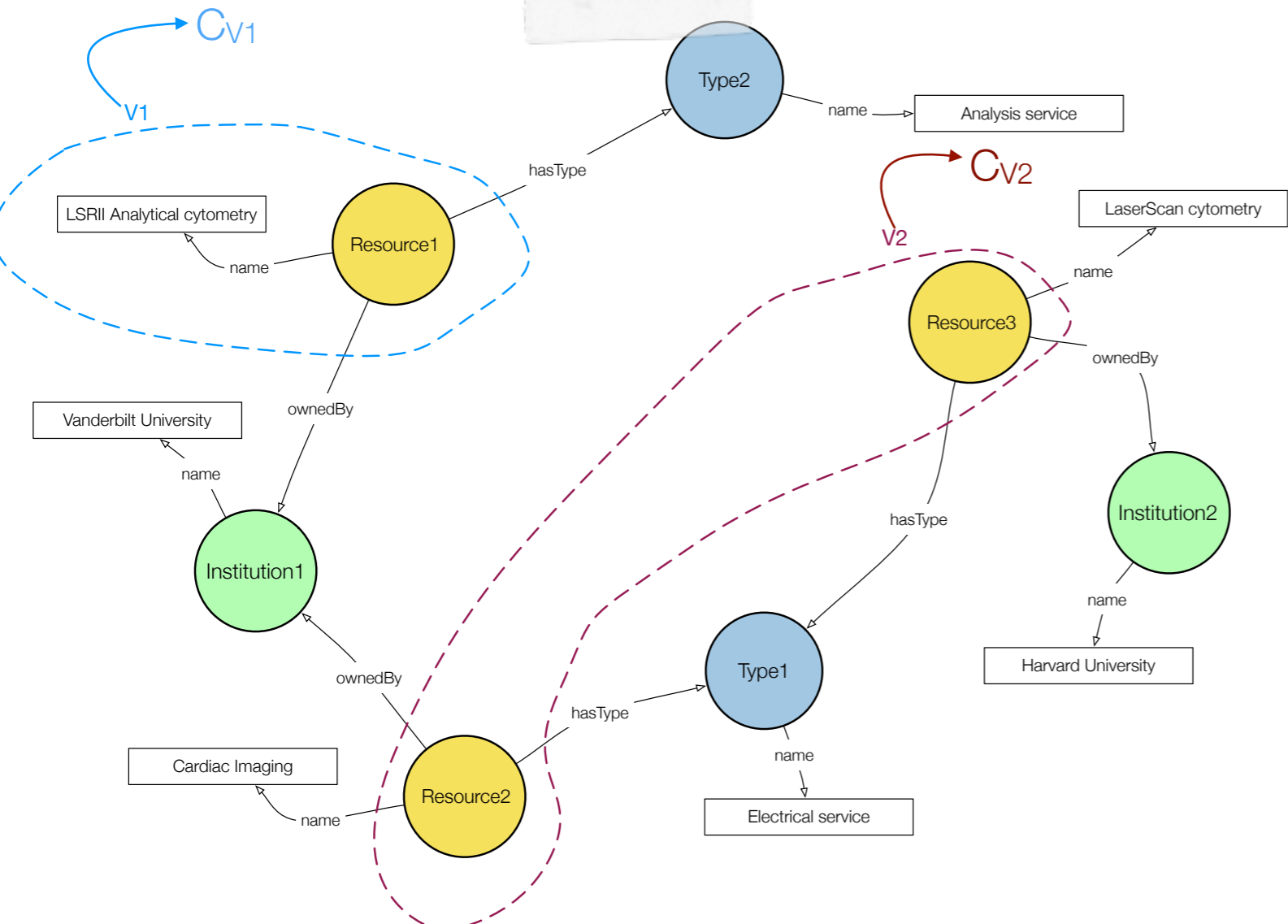
Creating a citation



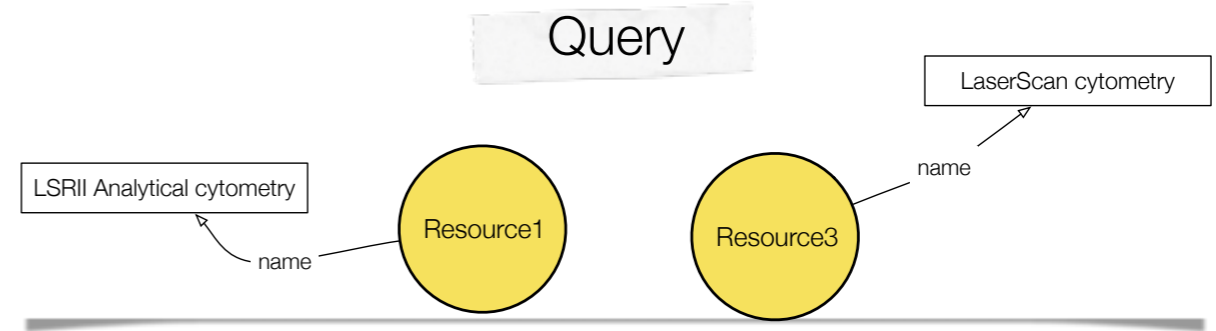
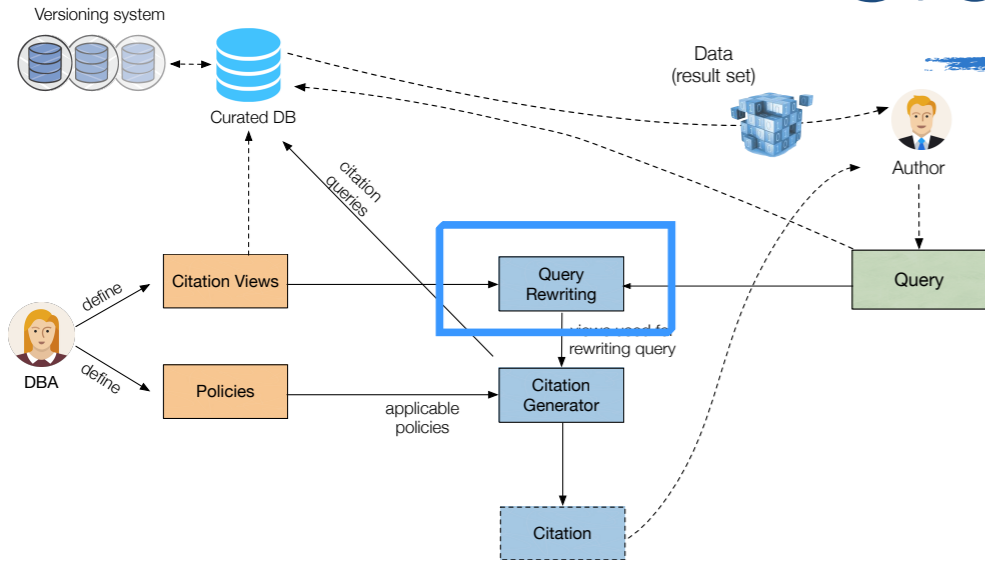
Creating a citation



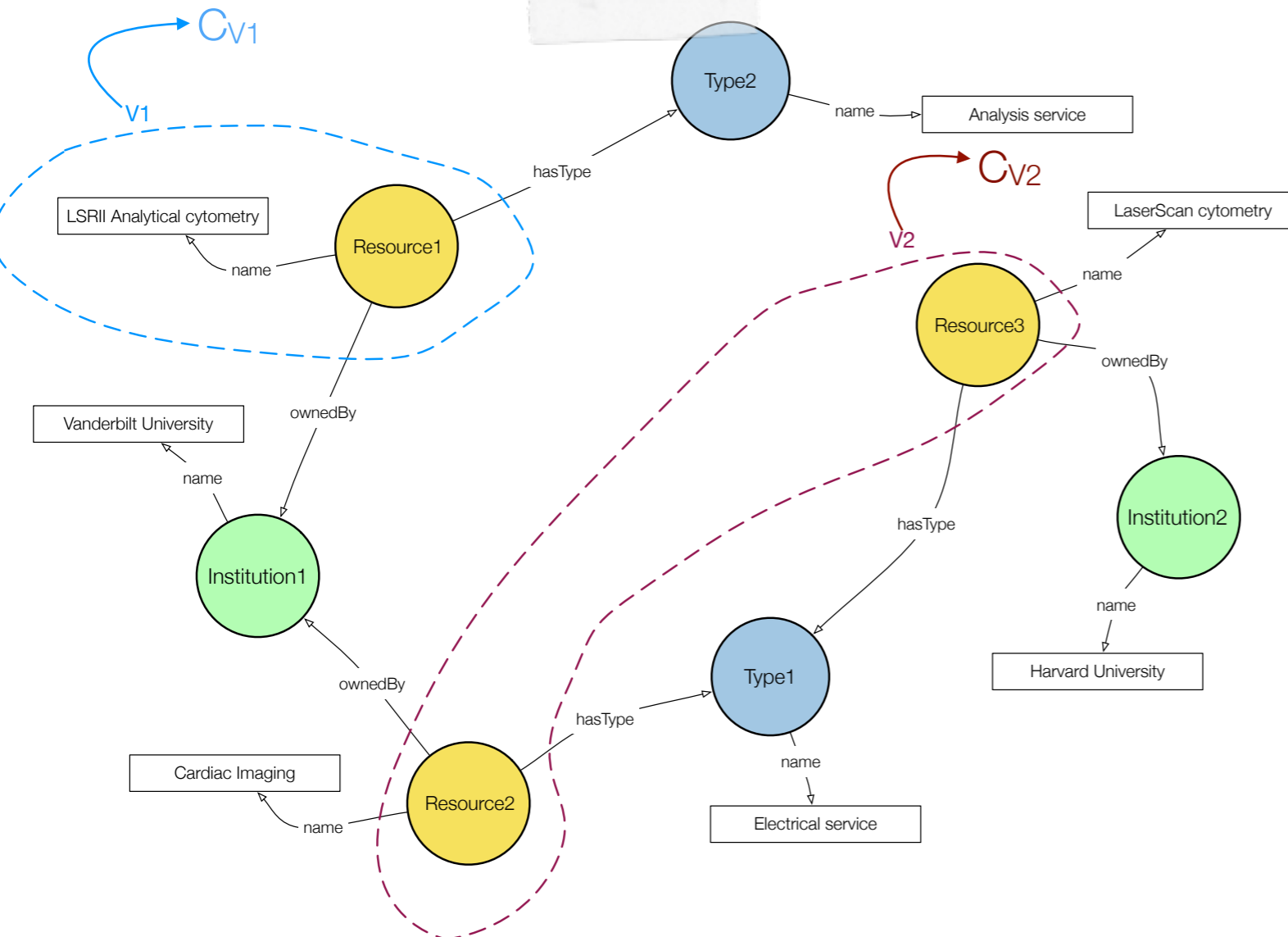
Definition of the citation views



Creating a citation

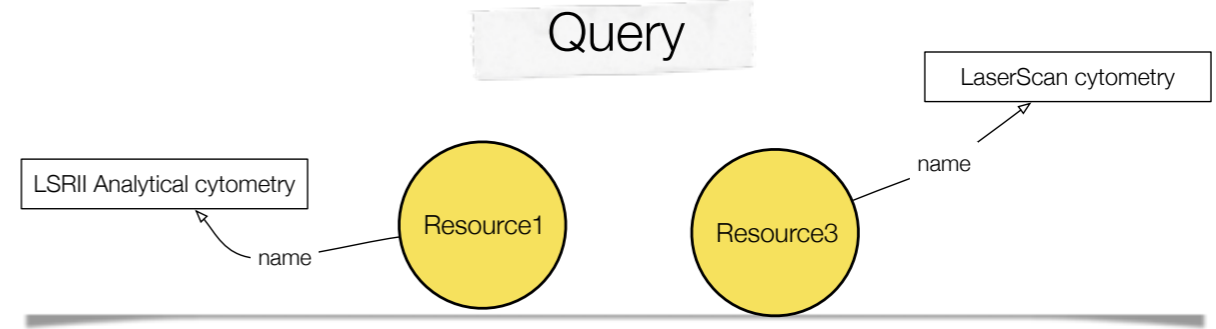
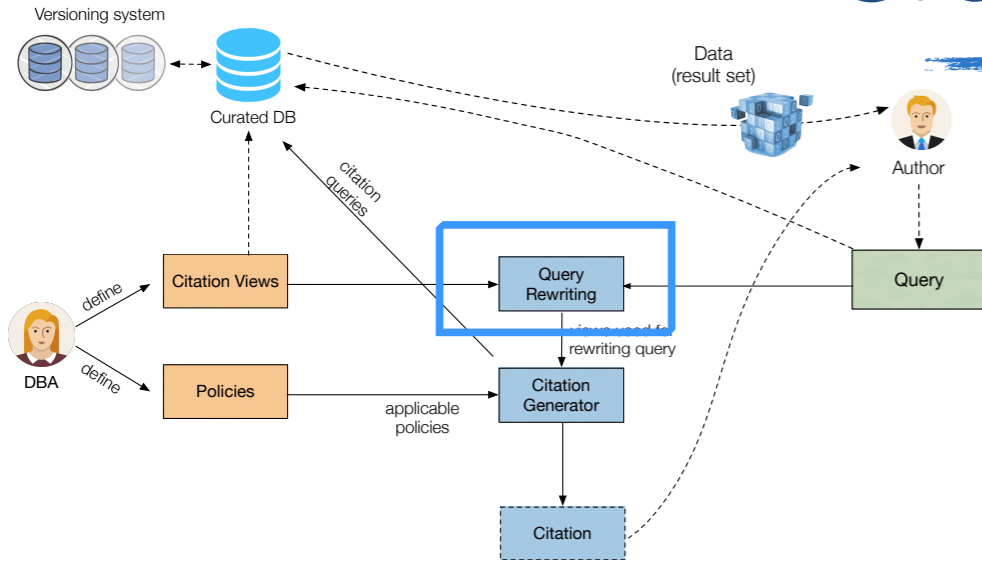


Definition of the citation views

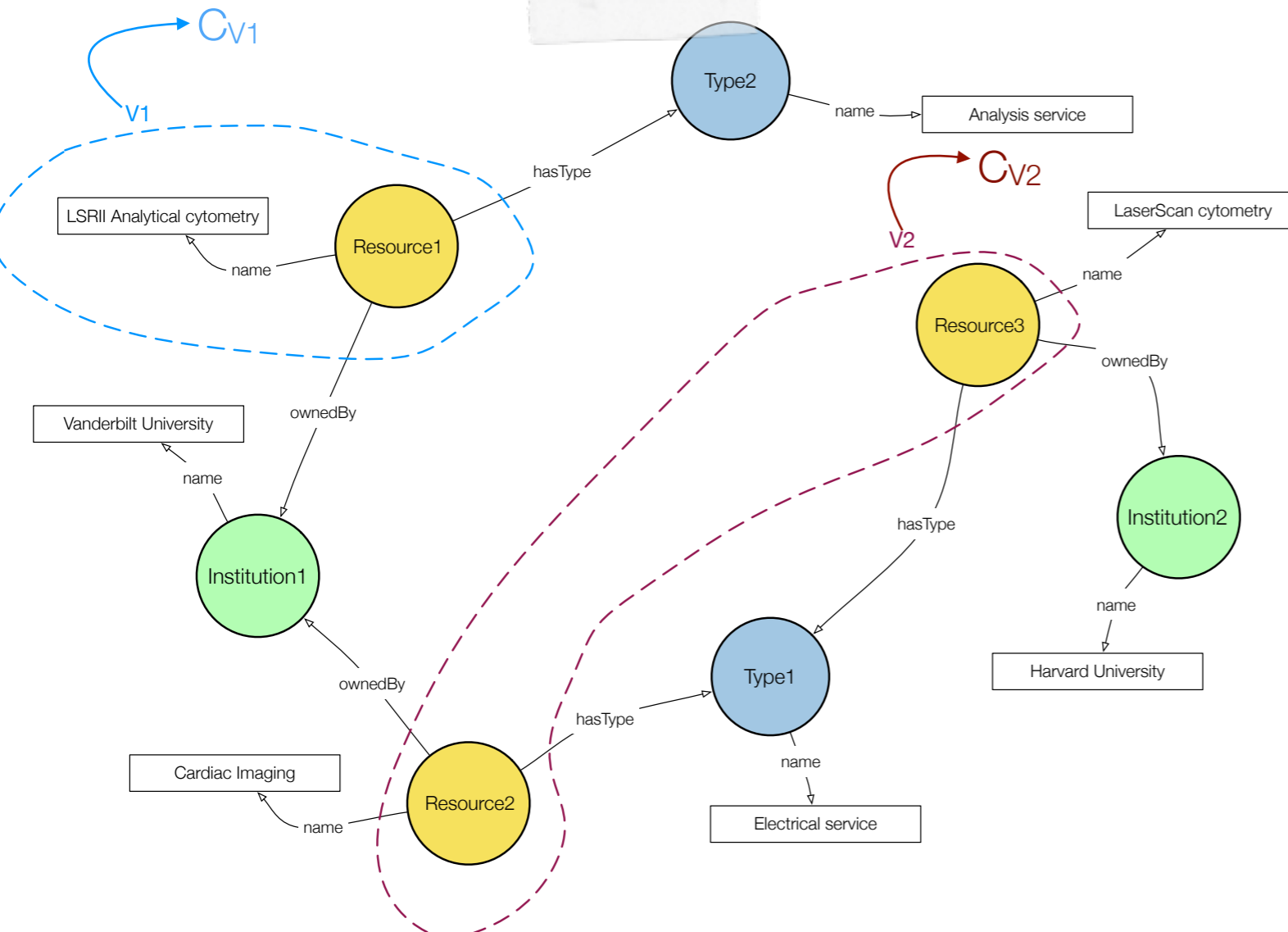


The query is rewritten using the views and the citation texts are created

Creating a citation



Definition of the citation views

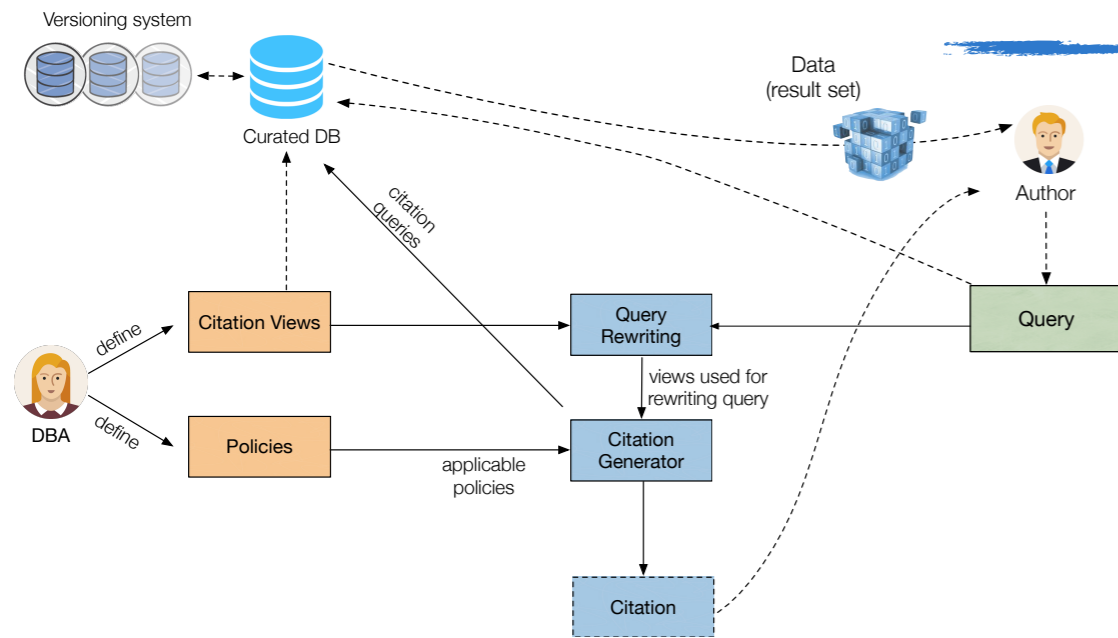


The query is rewritten using the views and the citation texts are created

V1(Resource1) -> C_{V1}(Resource1) -> {Grant, G., Version12, LSRII Analytical cytometry, doi.10/resource1}

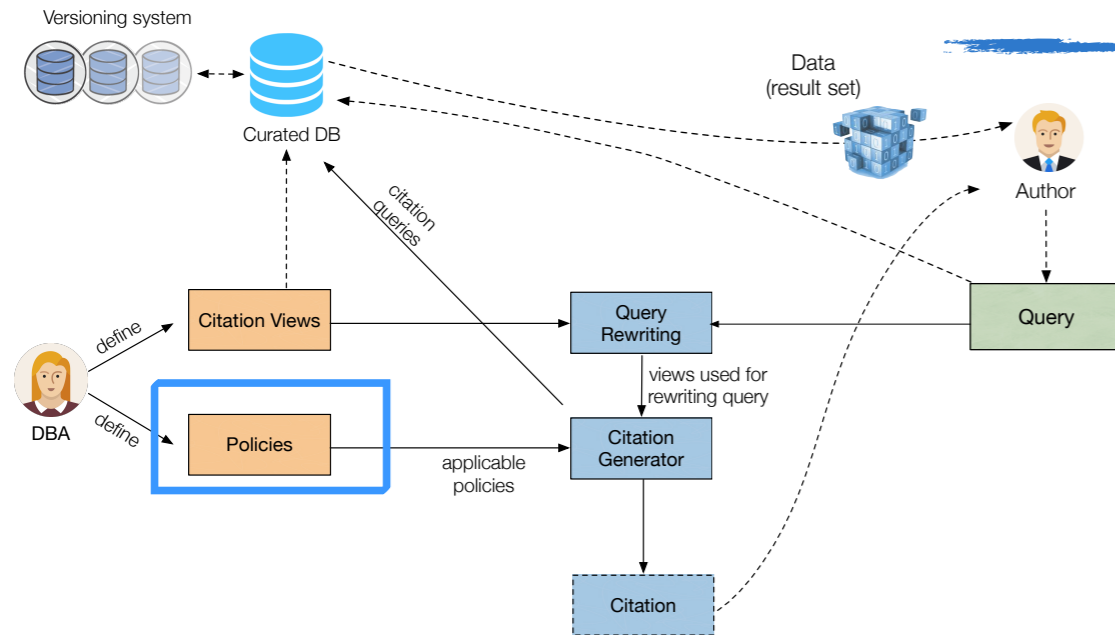
V2(Resource3) -> C_{V2}(Resource3) -> {Grant, G., Version12, LaserScan cytometry, doi.10/resource3}

Creating a citation



The citation policies are used to define how to create a single citation text

Creating a citation

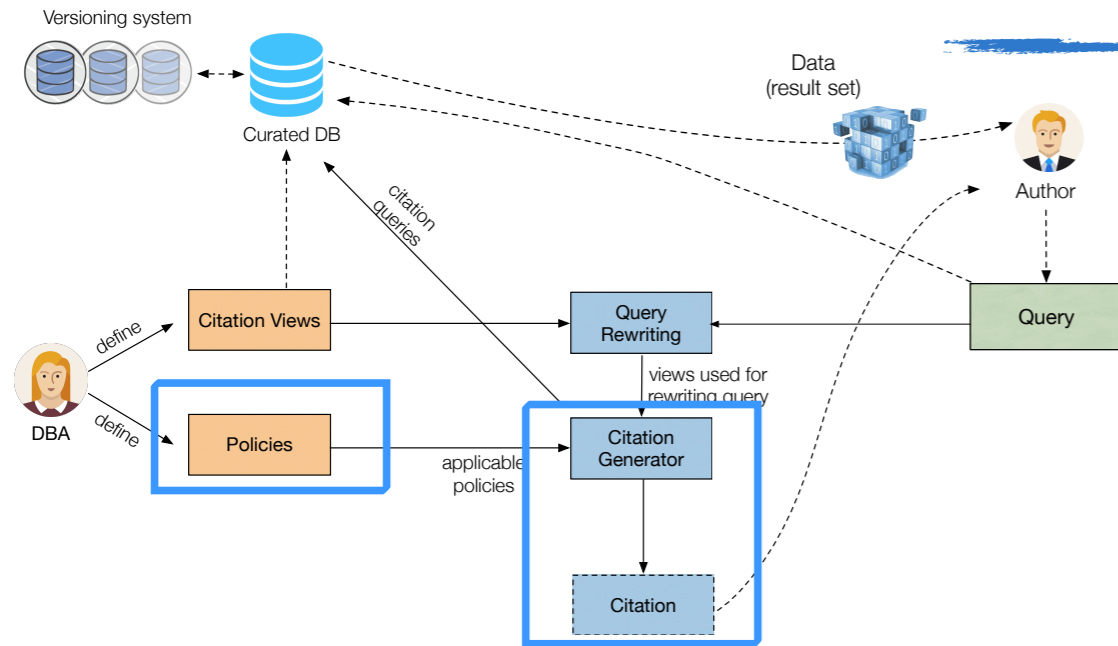


The citation policies are used to define how to create a single citation text

V1(Resource1) ->
{Grant, G., Version12,
LSRll Analytical cytometry, doi.10/resource1}

V2(Resource3) ->
{Grant, G., Version12,
LaserScan cytometry, doi.10/resource3}

Creating a citation



The citation policies are used to define how to create a single citation text

V1(Resource1) ->
{Grant, G., Version12,
LSRII Analytical cytometry, doi.10/resource1}

V2(Resource3) ->
{Grant, G., Version12,
LaserScan cytometry, doi.10/resource3}

A possible citation policy is to take the union of the citations

{Grant, G., Version12.
LSRII Analytical cytometry, doi.10/resource1;
LaserScan cytometry, doi.10/resource3}

What's required to the DBA

- Understand what information must be captured in the database to populate the citations
- Specify the citation views for the database
- Specify the citation policies
- Ensure that the system is versioned and enable dereferencing

<http://wazirul.blogspot.it/2015/12/5-orang-yang-mendadak-jenius-setelah.html>



XML: Learning to Cite Framework

Application: Digital Archives

[Check JASIST2017a](#)

Use case: Digital archives

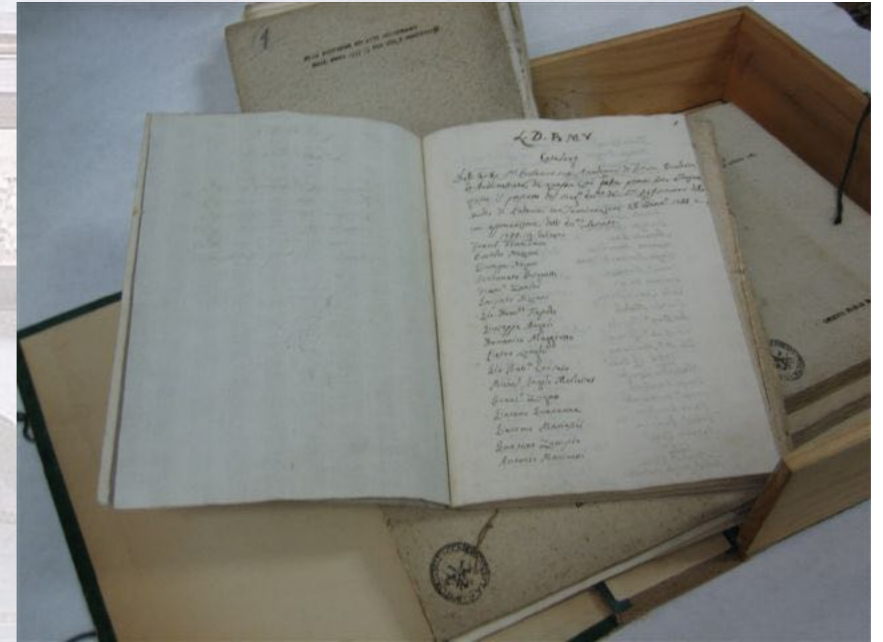
What is an Archive?



Shelves



Folders



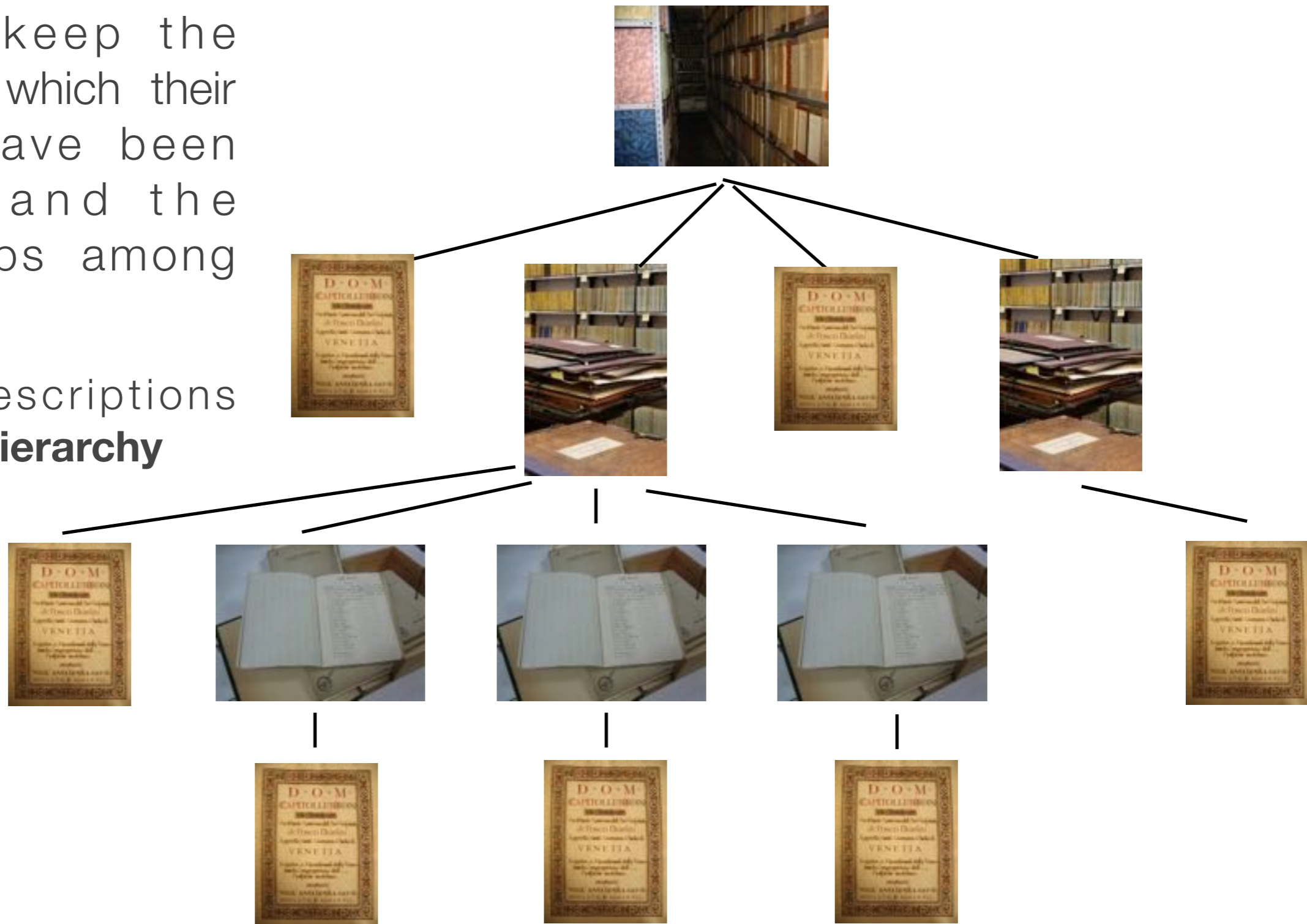
Envelops







Documents (e.g. letters, registers, testaments)

Archival Tree

- Archives keep the **context** in which their records have been created and the relationships among them
- Archival descriptions constitute a **hierarchy**

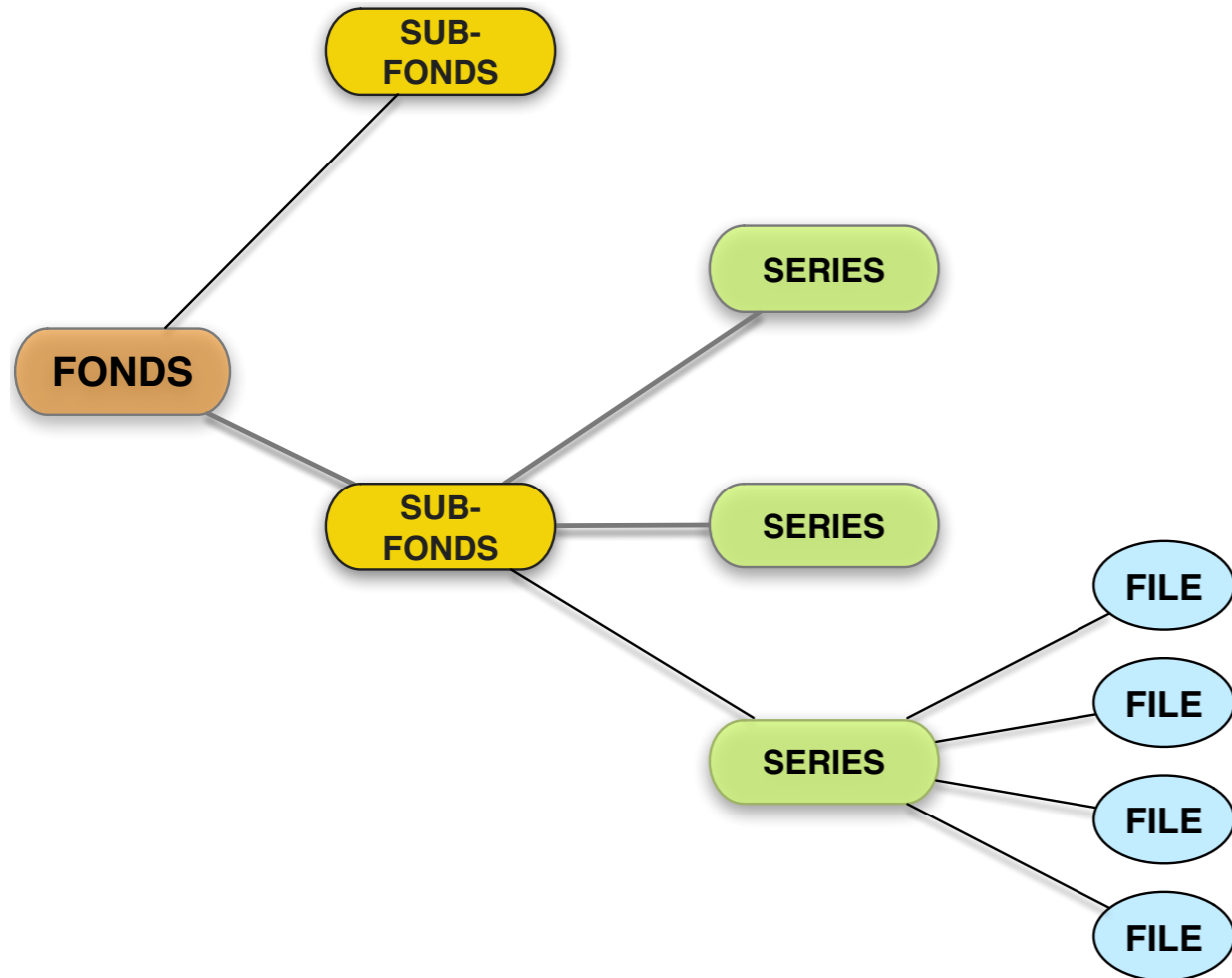


Archival Tree

-  **fonds**
-  **sub-fonds**
-  **series**
-  **document**



Encoding of archival data: EAD



(a) Archival Tree

```
<ead>
  <eadheader>
    [...]
  </eadheader>
  <archdesc level="fonds">
    [...]
    <did>[...]</did>
    <dsc level="fonds">
      [...]
      <c01 level="sub-fonds">
        [...]
      </c01>
      <c01 level="sub-fonds">
        [...]
        <c02 level="series">
          [...]
        </c02>
        <c02 level="series">
          [...]
        </c02>
        <c02 level="series">
          [...]
          <c03 level="file">
            [...]
          </c03>
          <c03 level="file">
            [...]
          </c03>
          <c03 level="file">
            [...]
          </c03>
          <c03 level="file">
            [...]
          </c03>
        </c02>
      </c01>
    </dsc>
  </archdesc>
</ead>
```

(b) EAD representation

Characteristics of EAD files

- A single EAD file encodes a whole archive
- “Big” XML files with deep hierarchy
- Heterogeneous use of tags across collections and within the same collection
- Often textual elements contain HTML tags
- Every element and attribute of an EAD file is a potential citable unit

EAD: Some statistics

| Collection | Files | Nodes | | Depth | | Size (KB) | | Max Fan Out | |
|------------|-------|---------|--------|-------|--------|-----------|--------|-------------|--------|
| | | max | median | max | median | max | median | max | median |
| AH 2005 | 233 | 14,648 | 158 | 21 | 6 | 760 | 15 | 1,332 | 23 |
| IISG 2005 | 798 | 52,213 | 513 | 17 | 9 | 2,290 | 34 | 2,601 | 21 |
| NA 2008 | 1681 | 160,061 | 880.5 | 18 | 9 | 9,750 | 58 | 10,271 | 34 |
| LoC 2014 | 2083 | 188,862 | 685 | 18 | 10 | 15,510 | 58 | 5,000 | 32 |
| UniMa 2014 | 662 | 69,766 | 711 | 10 | 8 | 2,960 | 40 | 6,861 | 43 |

AH 2005: UK Archival Hub, 2005 snapshot

IISG 2005: International Institute of Social History, 2005 snapshot

NA 2008: Nationaal Archief, The Netherlands, 2008 snapshot

LoC 2014: Library of Congress, 2014 snapshot

UniMa 2014: University of Maryland, 2014 snapshot

A Human-readable citation

Citable unit

Correspondence, 1951-1956

Contextual Information (from ancestors of the citable unit)

"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:
Writings (1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.

<http://hdl.loc.gov/loc.mss/eadmss.ms001024>

(Persistent) Unique identifier of the EAD file

All the elements of the citations are obtained from the EAD file containing the citable unit

In general, EAD files always contain all the information required to build a citation and a citable unit alone cannot be used to create a complete citation

A machine-readable citation

Conjunction of XPaths

```
/ead/eadheader/eadid && /ead/eadheader/filedesc/publicationstmt/publisher && /ead/archdesc/did/unittitle && /ead/archdesc/dsc/c01[10]/did/unittitle && /ead/archdesc/dsc/c01[10]/did/unittitle/unitdate && /ead/archdesc/dsc/c01[10]/did/container/@type && /ead/archdesc/dsc/c01[10]/did/container && /ead/archdesc/dsc/c01[10]/c02/did/container/@type && /ead/archdesc/dsc/c01[10]/c02/did/container && /ead/archdesc/dsc/c01[10]/c02/did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/container/@type && /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/container && /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/c05[1]/did/unittitle
```

A machine-readable citation

Human-Readable Citation

Machine-Readable Citation

| | |
|---|---|
| http://hdl.loc.gov/loc.mss/eadmss.ms001024 ← | /ead/eadheader/eadid |
| Manuscript Division, Library of Congress ← | /ead/eadheader/filedesc/publicationstmt/publisher |
| Huntington Cairns Papers ← | /ead/archdesc/did/unittitle |
| Part II: Writings ← | /ead/archdesc/dsc/c01[10]/did/unittitle |
| 1905-1984 ← | /ead/archdesc/dsc/c01[10]/did/unittitle/unitdate |
| box ← | /ead/archdesc/dsc/c01[10]/did/container/@type |
| 129-152 ← | /ead/archdesc/dsc/c01[10]/did/container |
| By Cairns ← | /ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle |
| box ← | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type |
| 129 ← | /ead/archdesc/dsc/c01[10]/c02[1]/did/container/ |
| Books ← | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle |
| box ← | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type |
| 135 ← | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container |
| "The Elements of Legal Theory" (unpublished) ← | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle |
| Correspondence, 1951-1956 ← | /ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle |

What does the user see?

LIBRARY OF CONGRESS ASK A LIBRARIAN DIGITAL COLLECTIONS LIBRARY CATALOGS

Correspondence, 1951-19 1 di 1

The Library of Congress > Researchers > Search Finding Aids > Huntington Cairns papers, 1780-1984

Huntington Cairns papers, 1780-1984

Search this Finding Aid all words Search Search All Finding Aids Help Contact Us

Overview Contents List Index Terms Using this Collection Search Results Print/Download

< Previous Page | Next Page > Part II: Writings, 1905-1984 [>> Navigate Contents List](#)

Biography of Cairns's writings, circa 1965
Book reviews
1925-1945
(14 folders)
1946-1963, undated
(15 folders)

BOX 134

BOX 135

Books
The Collected Dialogues of Plato (1961)
Miscellany, 1961-1983, undated
Scrapbook, 1961-1964, undated
(3 folders)
"The Elements of Legal Theory" (unpublished)
Correspondence, 1951-1956
Draft, 1954-1958, undated
(2 folders)
Outline, undated
Goethe, Johann Wolfgang von, *Faust* (unpublished translation)
Correspondence, 1947-1953
(2 folders)
Translation drafts, 1947-1949, undated
(4 folders)
Great Paintings from the National Gallery of Art (1952), scrapbook, 1952-1954
H. L. Mencken: The American Scene, A Reader (1965)
Correspondence, 1965
Reviews, 1965
(2 folders)

BOX 136

Contents List

- Part I: General Correspondence, 1925-1964
- Part I: James Kern Feibleman File, 1938-1964
- Part I: Subject File, circa 1931-1944
- Part I: Book and Article File, circa 1926-1965
- Part I: Miscellany, 1962-1961
- Part II: Family Papers, 1816-1984
- Part II: General Correspondence, 1919-1984
- Part II: Subject File, 1920-1984
- Part II: Speeches, 1933-1973
- Part II: Writings, 1905-1984**
- Part II: Miscellany, 1780-1984
- Part II: Oversize, 1816-1977

Correspondence, 1951-1956,
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings
(1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.
<http://hdl.loc.gov/loc.mss/eadmss.ms001024>

What does the user see?

Huntington Cairns papers, 1780-1984

Search this Finding Aid all words [Search All Finding Aids](#) [Help](#) [Contact Us](#)

Overview Contents List Index Terms Using this Collection Search Results Print/Download

[Title Page](#) | [Collection Summary](#) | [Biographical/Organizational Note](#) | [Scope and Contents](#) | [Arrangement](#)

⚠ Some or all content stored offsite.

Collection Summary

| | |
|------------------------------|---|
| Title | Huntington Cairns papers, 1780-1984 |
| Span Dates | 1780-1984 |
| Bulk Dates | (bulk 1925-1984) |
| ID No. | MSS14746 |
| Creator | Cairns, Huntington, 1904-1985 |
| Extent | 53,450 items ; 167 containers plus 13 oversize ; 73.1 linear feet |
| Language | Collection materials in English |
| Location | Manuscript Division, Library of Congress, Washington, D.C. |
| Summary | A former government official, and lawyer. Correspondence, manuscripts and galley proofs of writings, speeches, subject and research files, family papers, printed material, scrapbooks, and other papers concerning Cairns's career with the U.S. Bureau of Customs as a federal censor of imported books and films, as a lawyer with the Maryland Tax Revision Commission (1938-1941), and as a writer on the arts, law, literature, and philosophy. |
| Finding Aid Permalink | Cite or bookmark this finding aid as: http://hdl.loc.gov/loc.mss/eadmss.ms001024 |
| LCCN Permalink | LC Online Catalog record for this collection: https://lccn.loc.gov/mm79014746 |

Correspondence, 1951-1956,
"The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings
(1905-1984), box 129-152. Huntington Cairns Papers.
Manuscript Division, Library of Congress.
<http://hdl.loc.gov/loc.mss/eadmss.ms001024>

Generation of citations: The problem

Given:

data

```
/ead/archdesc/dsc/c01[10]/  
c02/c03[4]/c04[2]/c05[1]/  
did/unittitle
```

Query: XPath



Dataset: EAD file



generate

citation

Human-Readable Citation

```
http://hdl.loc.gov/loc.mss/eadmss.ms001024 ←-----  
Manuscript Division, Library of Congress ←-----  
Huntington Cairns Papers ←-----  
Part II: Writings ←-----  
1905-1984 ←-----  
box ←-----  
129-152 ←-----  
By Cairns ←-----  
box ←-----  
129 ←-----  
Books ←-----  
box ←-----  
135 ←-----  
"The Elements of Legal Theory" (unpublished) ←-----  
Correspondence, 1951-1956 ←-----
```

Machine-Readable Citation

```
/ead/eadheader/eadid  
/ead/eadheader/filedesc/publicationstmt/publisher  
/ead/archdesc/did/unittitle  
/ead/archdesc/dsc/c01[10]/did/unittitle  
/ead/archdesc/dsc/c01[10]/did/unittitle/unitdate  
/ead/archdesc/dsc/c01[10]/did/container/@type  
/ead/archdesc/dsc/c01[10]/did/container  
/ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle  
/ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type  
/ead/archdesc/dsc/c01[10]/c02[1]/did/container/  
/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle  
/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type  
/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container  
/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle  
/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle
```

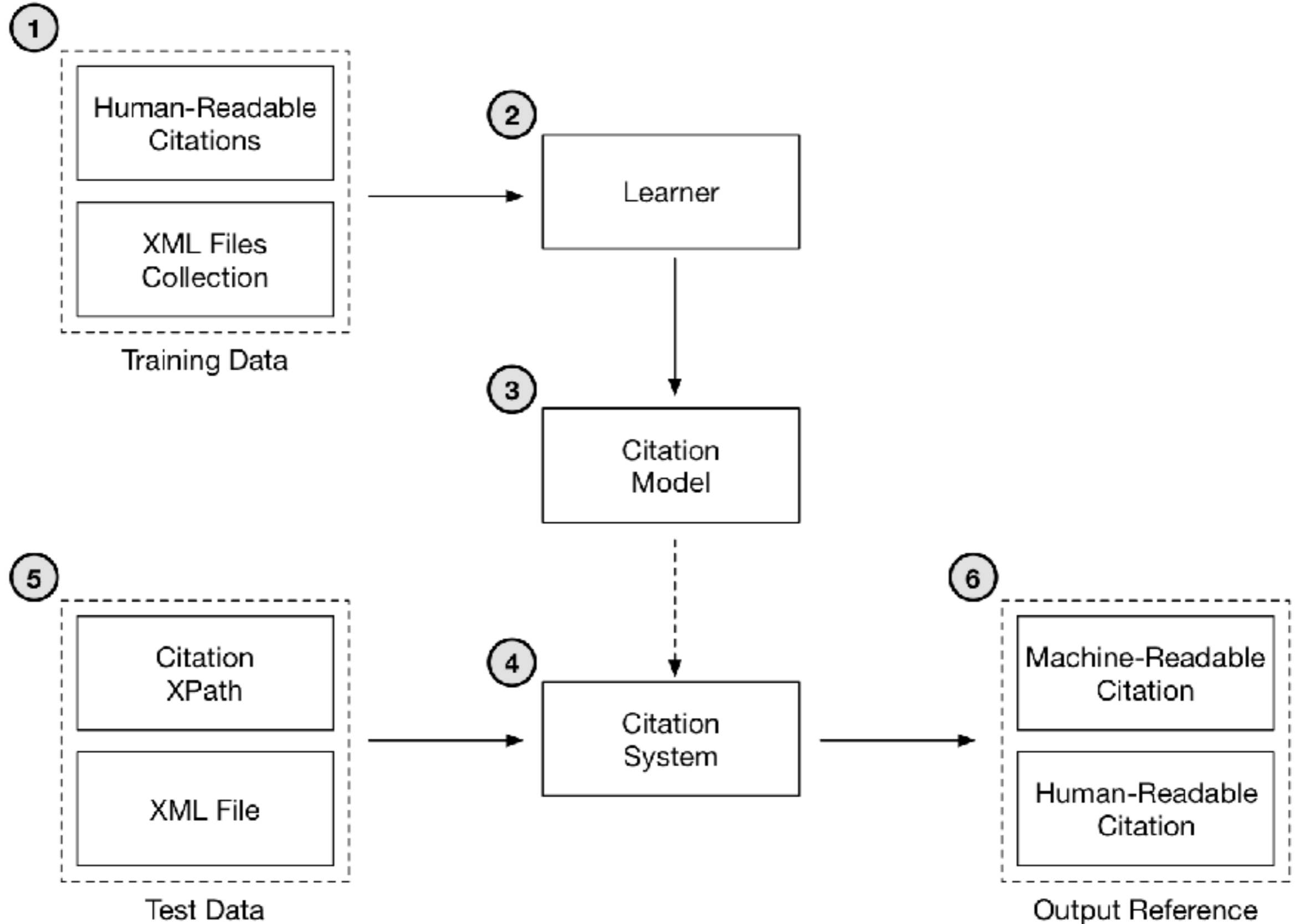
Learning to cite framework

Check [JASIST2017a](#)

Learning to cite framework

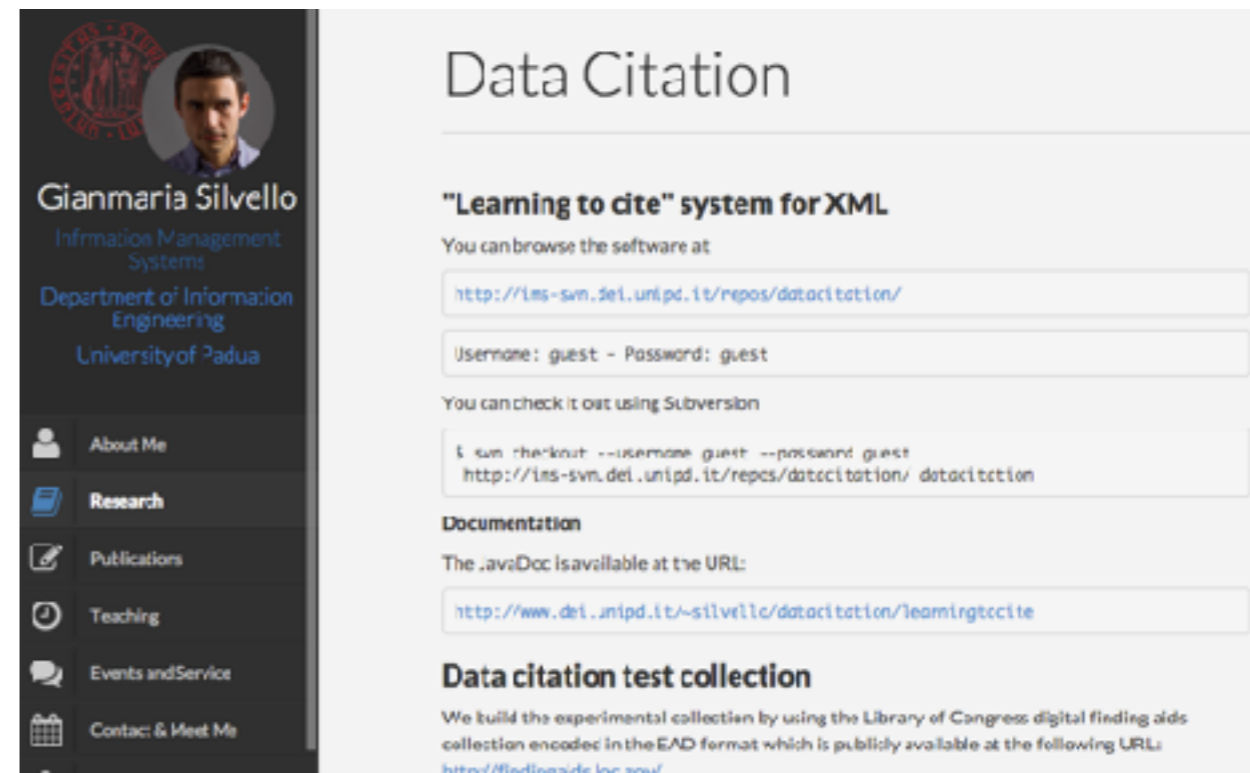
- The idea is to employ a machine learning approach for the generation of citations
- Learn from some sample data (human-readable citations), get a citation model out of it, and generate citations
- Require low effort (and resources) to data creators and curators
- Handle data heterogeneity

Learning to cite framework



System implementation and data

- The citation system is open-source and implemented in Java (Maven project) as well as the code for the experiments
- The training data, test data and the ground truth are openly available
- <http://www.dei.unipd.it/~silvello/datacitation/>



The screenshot shows the 'Data Citation' website. On the left is a dark sidebar with a profile for Gianmaria Silvello, including his name, affiliation (Information Management Systems, Department of Information Engineering, University of Padua), and a list of menu items: About Me, Research, Publications, Teaching, Events and Service, and Contact & Meet Me. The main content area is light gray and features the title 'Data Citation' at the top. Below it is a section titled '"Learning to cite" system for XML' with a sub-header 'You can browse the software at' and a text input field containing the URL `http://ims-svn.dei.unipd.it/repos/datacitation/`. A second input field shows 'Username: guest - Password: guest'. Below this is a section 'You can check it out using Subversion' with a code block: `svn checkout --username guest --password guest http://ims-svn.dei.unipd.it/repos/datacitation/ datacitation`. The 'Documentation' section states 'The .javaDoc is available at the URL:' followed by an input field with `http://www.dei.unipd.it/~silvello/datacitation/learningtocite`. The final section is 'Data citation test collection', which explains that the experimental collection is built using the Library of Congress digital finding aids collection encoded in the EAD format, publicly available at `http://findingaids.loc.gov/`.

Citing Relational DB Using Views

Application: IUPHAR/BPS

Check CIDR2017 + PODS2017

Joint work with Susan Davidson, Daniel Deutch and Tova Milo

Some slides are taken from Davidson's presentation @PODS2017 and Silvello's presentation @SEBD2017

G. Silvello - Automatically generating citation text from queries

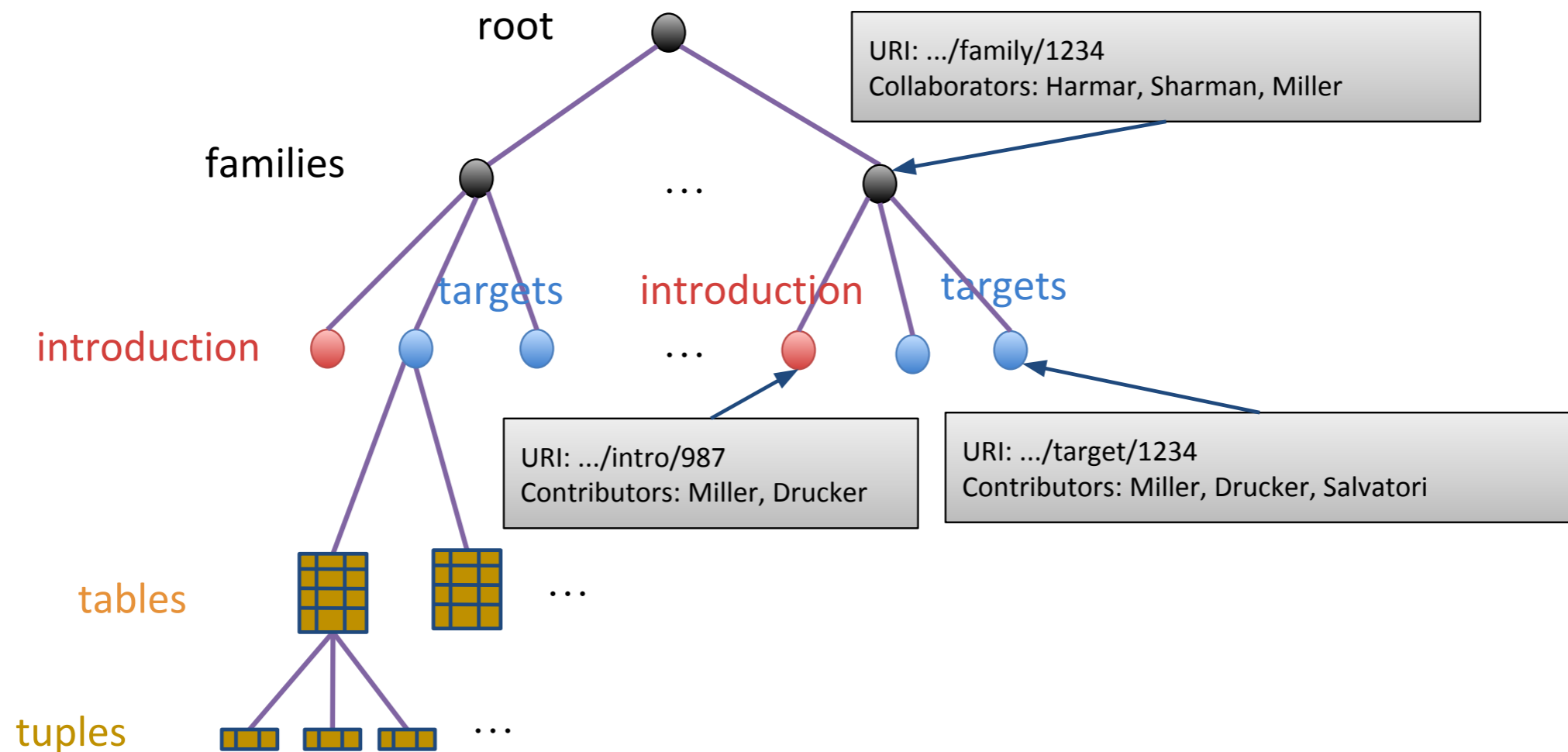
The citation generation problem

- It is common for database owners to supply citations for some parts (**views**) of the database, $V_1 \dots V_n$.
- So the problem becomes: Given a query Q , can it be rewritten using the views? That is, is there a Q_i such that
- $$\forall D \in S. Q(D) = Q_i(V_{i1}(D), \dots, V_{ik}(D))$$
- If so, the citations for $V_{i1} \dots, V_{ik}$ could be used to create a citation for Q .

Answering queries using views

- The problem of answering queries using views has been well studied and is generally hard – but in our context may be tractable.
- A. Halevy. Answering queries using views: A survey. *VLDB J.*, 10(4):270–294, 2001.
- Lenzerini. Data Integration: A Theoretical Perspective: *PODS*, 2002.
- A. Deutsch, L. Popa, and V. Tannen. Query reformulation with constraints. *SIGMOD Record*, 35(1):65–73, 2006.
- F. Afrati, C. Li and J. Ullman. Using views to generate efficient evaluation plans for queries. *JCSS* 73(5): 703 - 724, 2007.

“Parameterized” Views



Taken from Davidson's presentation @PODS2017

Effect of parameters

Family

| FID | FName | Type |
|-----|---|--------|
| 1 | Glucagon receptor... | GPCR |
| 2 | CLR (calcitonin receptor-like receptor) | GPCR |
| 3 | Peptidases and proteinases... | Kinase |
| 4 | A multifunctional molecule, | Kinase |
| 5 | Chromatin modifying enzymes... | Kinase |

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$

“Instantiated views”: $V1(F, N, Ty)(1), V1(F, N, Ty)(2), \dots, V1(F, N, Ty)(5)$

Effect of parameters

Family

| FID | FName | Type |
|-----|---|--------|
| 1 | Glucocorticoid receptor... | GPCR |
| 2 | CLR (calcitonin receptor-like receptor) | GPCR |
| 3 | Peptidases and proteinases... | Kinase |
| 4 | A multifunctional molecule, | Kinase |
| 5 | Chromatin modifying enzymes... | Kinase |

V4(F, N, Ty) :- Family(F, N, Ty)

Citation views

- To specify a citation, there are three components:
 - **View definition:** specifies what is being cited
 - **Citation query:** specifies what snippets of information to include in the citation
 - **Citation function:** specifies how to construct the citation from the snippets of information
- We call this triple a **citation view**.
- For now, we will focus on the view definition, which is expressed in Datalog.

One simple example

IUPHAR: Citation views

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)
Person(PID, PName, Affiliation)
FC(FID, PID) FIC (FID, PID)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

Citation queries:

$\lambda F. C_{V1}(F, PN) :- \text{Family}(F, N, Ty), \text{FC}(F, P), \text{Person}(P, PN)$
 $\lambda F. C_{V2}(F, PN) :- \text{FamilyIntro}(F, Tx), \text{FIC}(F, P), \text{Person}(P, PN)$

Generating citations

- If the query matches a view definition, we can use the associated citation query and function.
 - “Match” must be extended to take parameters into account.
- But what if it doesn't?
 - Nothing matches the query
 - A set of view definitions are used to rewrite the query
 - More than one set of view definitions can be used to rewrite the query

What is a “good” citation?

- Contains appropriate snippets of information
- Allows the data as it appeared at time of citation to be retrieved
- Concise
- Specific
- Our approach enables the DBA to specify the tradeoff between conciseness and specificity.

IUPHAR: Generating the citation (1)

- A query is another Datalog expression (unparameterized).

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

- This can be rewritten using V1

$Q_1(F, N, Ty) :- \text{Family}(F, N, Ty), F = 1$

- We can then construct a citation to Q in terms of the citation for V1(F, N, Ty)("1").

$Q_1'(F, N, Ty) :- V1(F, N, Ty)(1)$

Citation views as annotation

- Citation views are a type of annotation on tuples.
- Provenance is a form of annotation on tuples, which is well understood while being carried through queries.
 - Green, Karvounarakis, Tannen: Provenance Semirings, PODS 2007: 31-40.
 - **Joint use**: joins of tuples
 - **Alternate use**: unions and projections of tuples
- Can we use these ideas to understand how citation “annotations” on tuples are combined in general queries?

Citation “semiring”?

- Given a (conjunctive) query, we rewrite it to a set of minimal equivalent queries that contain at least one citation view.

- Let the set of queries obtained in this way be $\{Q_1, \dots, Q_n\}$

- Each Q_i contains a set of citation views $\{V_{i1}, \dots, V_{imi}\}$. The **joint** use (*) of their citations constructs a citation for Q_i , $C(Q_i)$.

- $$C(Q_i) = C(V_{i1})^* \dots^* C(V_{imi})$$

- The **alternate** use (+) of each $C(Q_i)$ constructs a citation for Q , $C(Q)$.

- $$C(Q) = C(Q_1) + \dots + C(Q_n)$$

*“Model for Fine-Grained Data Citation”, CIDR 2017
S. Davidson, D. Deutch, T. Milo, and G. Silvello.*

Interpreting * and +

- **Joint** use of citations: $C(V_{i1}) * \dots * C(V_{imi})$
 - * could be union or some sort of join
 - E.g. in example 4, V1 and V2 were jointly used: $V1(F, N, Ty)$
 $(“F123”) * V2(F, Tx)(“F123”)$
- **Alternate** use of citations: $C(Q_1) + \dots + C(Q_n)$
 - + could be union or min (wrt some ordering on views)
 - E.g. in example 3, both the parameterized and unparameterized views on Family matched $(V1(F, N, Ty)(1), V1(F, N, Ty)(2), \dots, V1(F, N, Ty)(5)) + V4$
- **Joint and alternate use are “policies” specified by the DBA**

Example of output citation

View definition:

$\lambda F. V1(F, N, Ty) :- Family(F, N, Ty)$

Citation query:

$\lambda F. C_{V1}(F, PN) :- Family(F, N, Ty), FC(F, P), Person(P, PN)$

$Q_1(F, N, Ty) :- Family(F, N, Ty), F= 1$

$Q_1'(F, N, Ty) :- V1(F, N, Ty)(1)$

Citation:

Miller, Drucker, Bataille, Chan, Delagrangue, Göke, Mayo, Thorens, Hills.
Glucagon receptor family.
Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
Family(F, N, Ty), F= 1

| FID | FName | Type |
|-----|--------------|------|
| 1 | Glucagen ... | GPCR |

Example, with * as “join”

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$

$\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

Citation queries:

$\lambda F. C_{V1}(F, PN) :- \text{Family}(F, N, Ty), \text{FC}(F, P), \text{Person}(P, PN)$

$\lambda F. C_{V2}(F, PN) :- \text{FamilyIntro}(F, Tx), \text{FIC}(F, P), \text{Person}(P, PN)$

$Q_1(F, N, Ty, Tx) :- \text{Family}(F, N, Ty), \text{FamilyIntro}(F, Tx), F=1$

$Q_1'(F, N, Ty, Tx) :- V1(F, N, Ty)(1), V2(F, Tx)(1)$

Citation:

Miller, Drucker, Bataille, Chan, Delagrangé,
Göke, Mayo, Thorens, Hills.
Glucagon receptor family, introduction.

Miller, Drucker, Bataille, Chan, Delagrangé,
Göke, Mayo, Thorens, Hills.
Glucagon receptor family.

Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
Family(F, N, Ty), FamilyIntro(F, Tx), F=1

| FID | FName | Type | Text |
|-----|--------------|------|------------------------|
| 1 | Glucagen ... | GPCR | Glucagon regulates ... |

The big picture

- **Database owners** need to be able to specify citation views for the database – schema level information.
- **Database users** (“authors”) need to have citations “served up” as they extract data through queries.
- **Dereferencing** the citation should bring back the data to “readers” as of the time it was cited.

Computational challenges

- Schema-level versus instance level?
 - Should we store the citations as annotations on tuples, or should we reason at the schema level and then calculate the citation?
- Given an expected query workload, what are the “best” citation views?
 - And are the necessary snippets of citation information in the schema?
- The number of alternative uses of citation views can be large.
 - Are there efficient algorithms to find the “best” according to some metric of quality (e.g. involving the number of views, the specificity of views, or related to a view hierarchy)?

Final remarks

- In general, the automatic generation of citation snippet is overlooked by many scientific databases
- Automatic generation of citation text is a key aspect of data citation
- Citation systems need to be customised for specific domains
- We are looking for interesting and challenging datasets to work with (new domains, requirements, users, ...)

References

- [JASIST2017a] G. Silvello (2017). Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data, *Journal of the Association for Information Science and Technology (JASIST)*, vol. 63 num. 6, pp. 1505–1524, June 2017. DOI: 10.1002/asi.23774
- [JASIST2017b] G. Silvello (2017). Theory and Practice of Data Citation, *Journal of the Association for Information Science and Technology (JASIST)*, available as early view, pages 1–24, 2017. DOI: 10.1002/asi.23917
- [CIDR2017] S. B. Davidson, D. Deutch, T. Milo and G. Silvello (2017). A Model for Fine-Grained Data Citation. In *Proc. of CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research*, 2017.
- [PODS2017] S. B. Davidson, P. Buneman, D. Deutch, T. Milo and G. Silvello (2017). Data Citation: a Computational Challenge. In *Proc. of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2017)*, pages 1–4, ACM Press, New York, NY, USA, 2017.
- [JCDL2017] A. Alawini, L. Chen, S. B. Davidson, N. Portilho Da Silva and G. Silvello (2017). Automating data citation: the eagle-i experience. In *Proc. of the 2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017*, pages 169–178, IEEE Computer Society, 2017.
- [IEEE2016] G. Silvello and N. Ferro (2016). Data Citation is Coming. Introduction to the Special Issue on Data Citation, *Bulletin of IEEE Technical Committee on Digital Libraries*, Volume 12 Issue 1, pp. 1–5, May 2016.
- [Dlib2015] G. Silvello (2015). A Methodology for Citing Linked Open Data Subsets, *D-Lib Magazine*, 21(1/2). DOI: 10.1045/january2015-silvello
- [IEEE2010] P. Buneman and G. Silvello (2010). A Rule-Based Citation System for Structured and Evolving Datasets. *Bulletin of the Technical Committee on Data Engineering Bulletin*, September 2010, 3(3):33–41. IEEE Computer Society. ISSN 1053-1238.

ANY
QUESTIONS
?