

Data Citation – Technical Issues - Identification



Herbert Van de Sompel
Los Alamos National Laboratory
Research Library

<http://public.lanl.gov/herbertv/>
hvdsomp@gmail.com
[@hvdsomp](#)



Identification of Data

- As part of a Citation
- To support various use case:
 - Credit
 - Access
 - Re-use
- To support different agents:
 - Human
 - Machine



Identification of Data as Part of Citation

- Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R. A
Principal Component Analysis of 39 Scientific Impact Measures.
PLoS ONE, 4(6), pp. e6022, 2009. **doi:10.1371/journal.pone/0006022**
- Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R. A
Principal Component Analysis of 39 Scientific Impact Measures.
PLoS ONE, 4(6), pp. e6022, 2009. **doi:10.1371/journal.pone/0006022** **<http://dx.doi.org/10.1371/journal.pone/0006022>**
- Bollen, J., Van de Sompel, H., Hagberg, A., Chute, R. A
Principal Component Analysis of 39 Scientific Impact Measures.
PLoS ONE, 4(6), pp. e6022, 2009. **<http://dx.doi.org/10.1371/journal.pone/0006022>**



Disclaimer

- Throughout, DOI is used as example of commonly used *Persistent Identifier*.
- Other commonly used Persistent Identifiers include ARK, HDL, URN:NBN
- I do not consider myself a dataset expert, but I have been running into identifiers for a long time

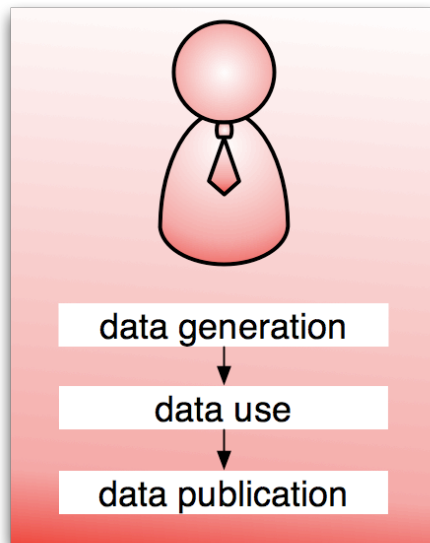


3 Considerations that Merit More Attention

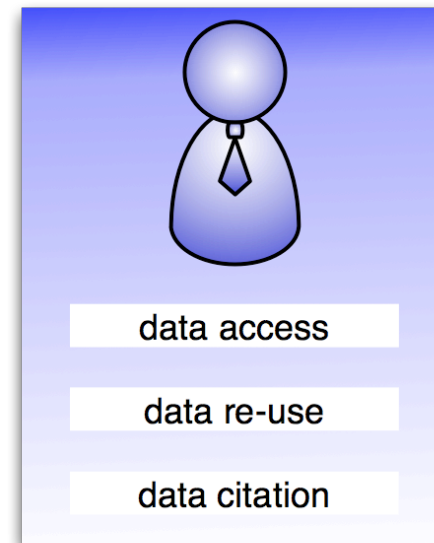
- Nature of Identifiers for Citation, Access, Re-Use
- Catering to Human and Machine Agents
- Granularity: Spatial ; Temporal



Data Publication, Access, Re-Use, Citation



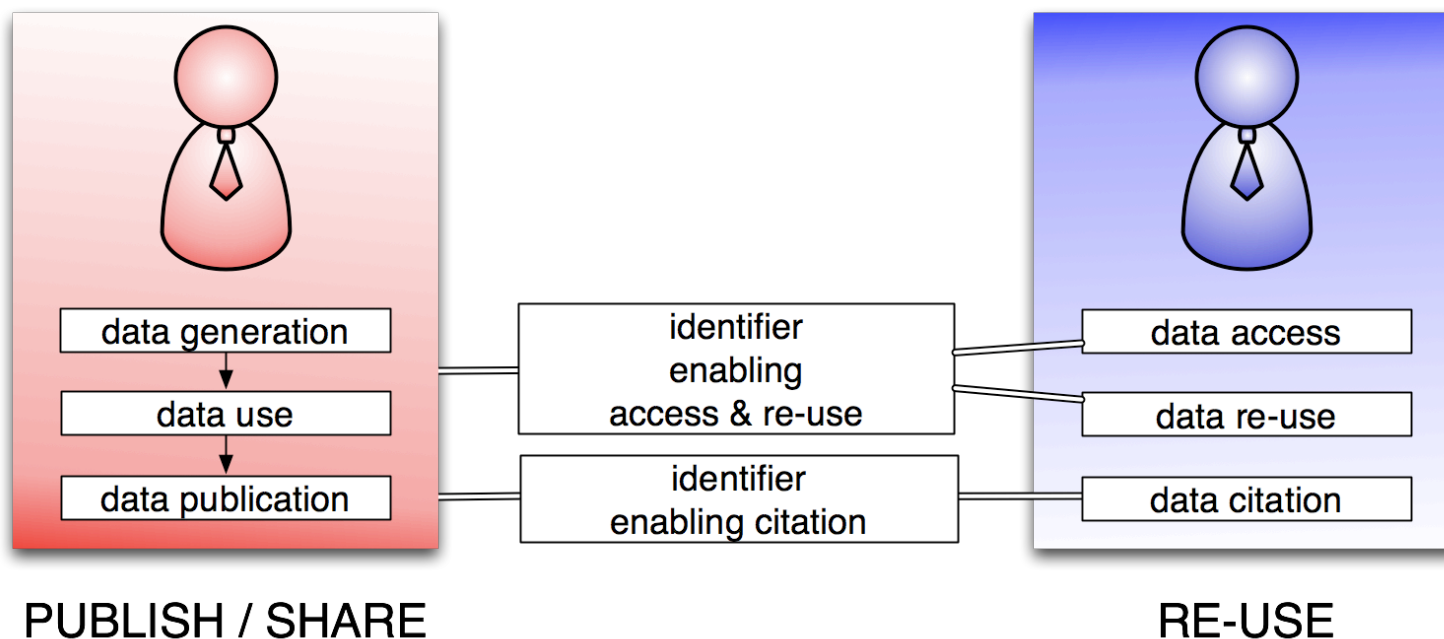
PUBLISH / SHARE



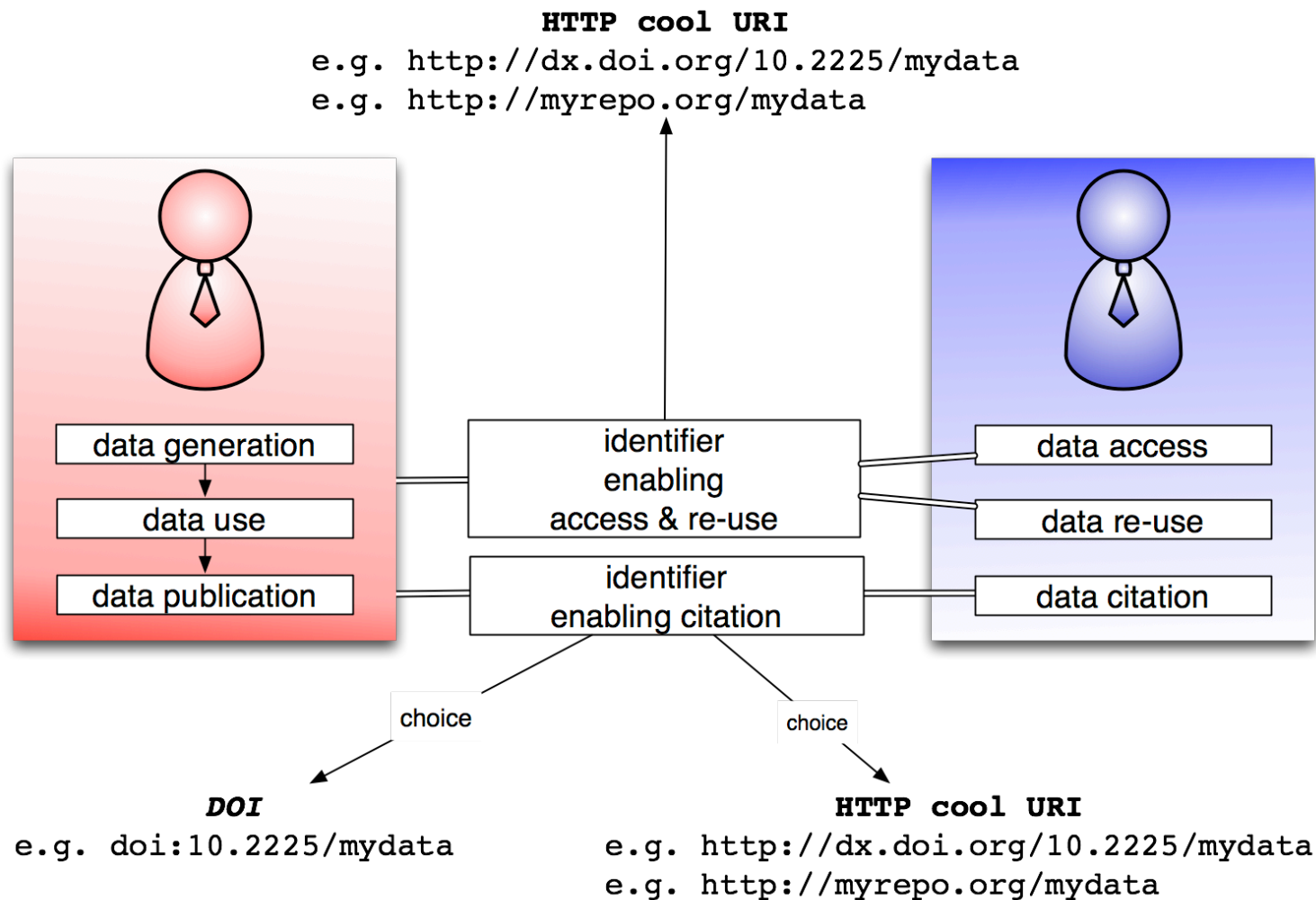
RE-USE



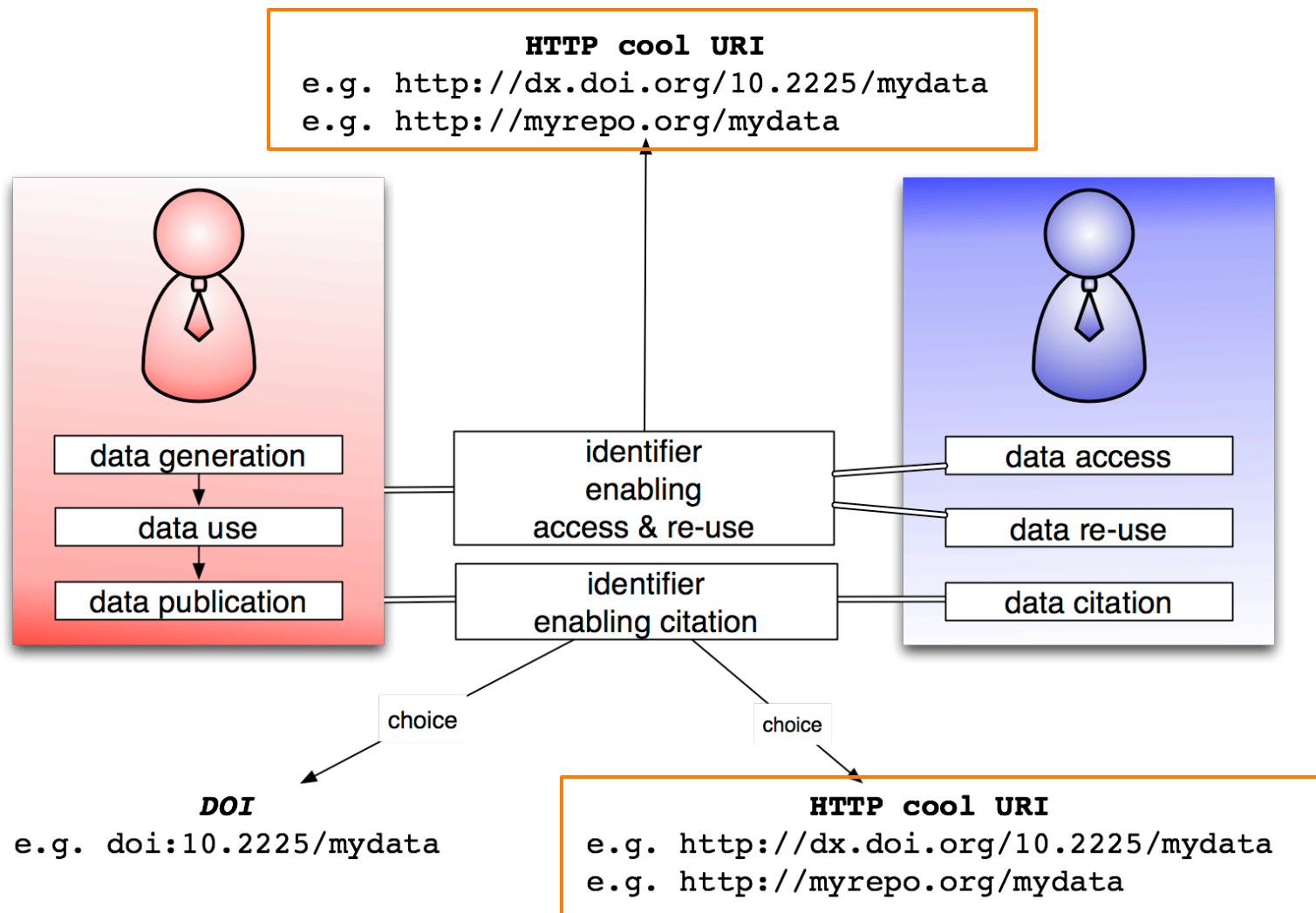
Consideration 1: Identifiers for Citation, Access, Re-Use



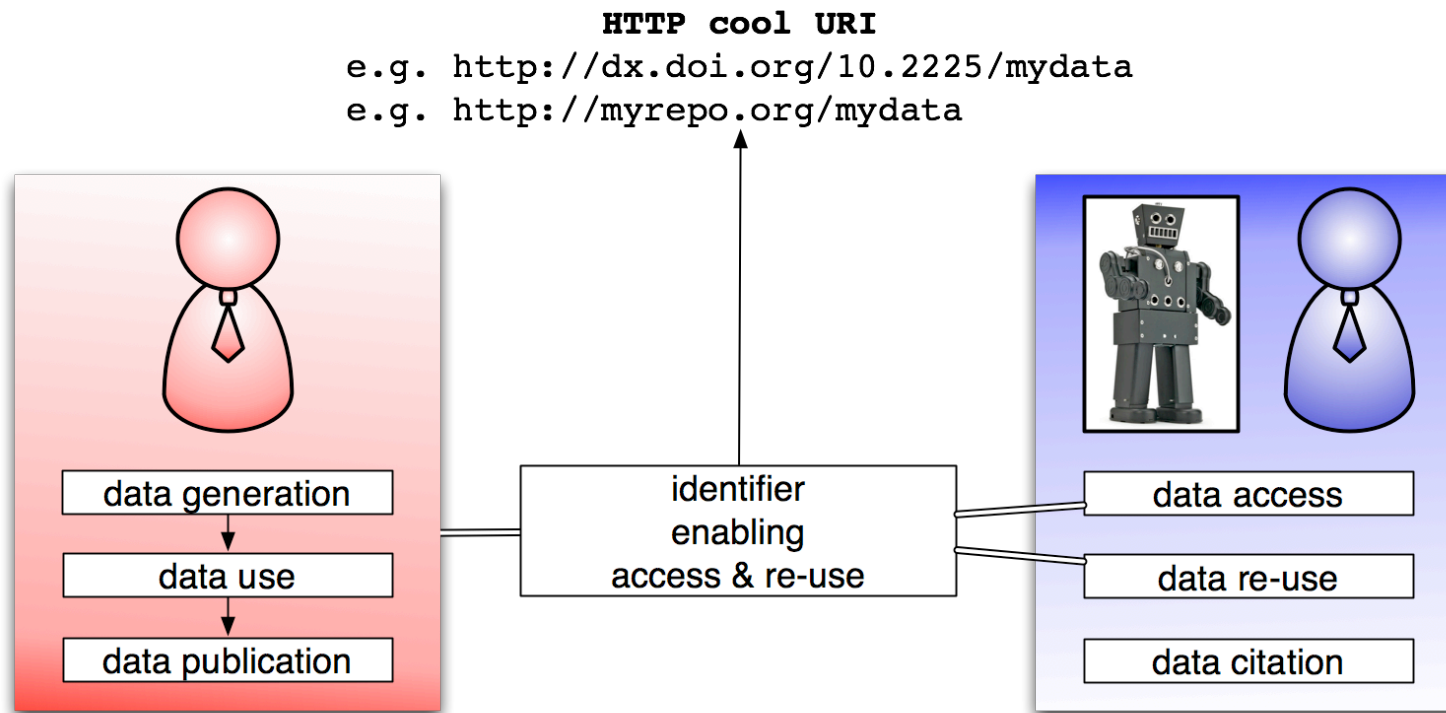
Identifiers for Citation, Access, Re-Use



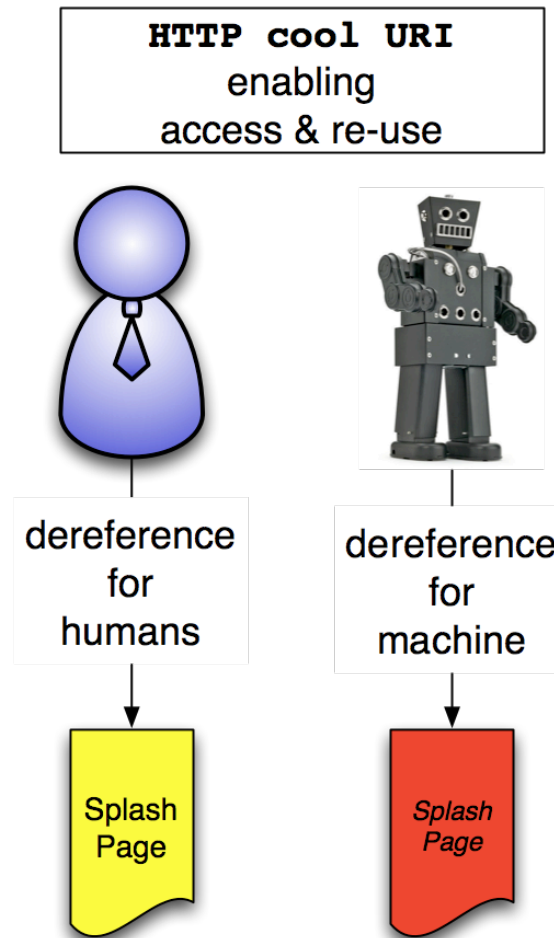
Identifiers for Citation, Access, Re-Use



Consideration 2: Catering to Human and Machine Agents



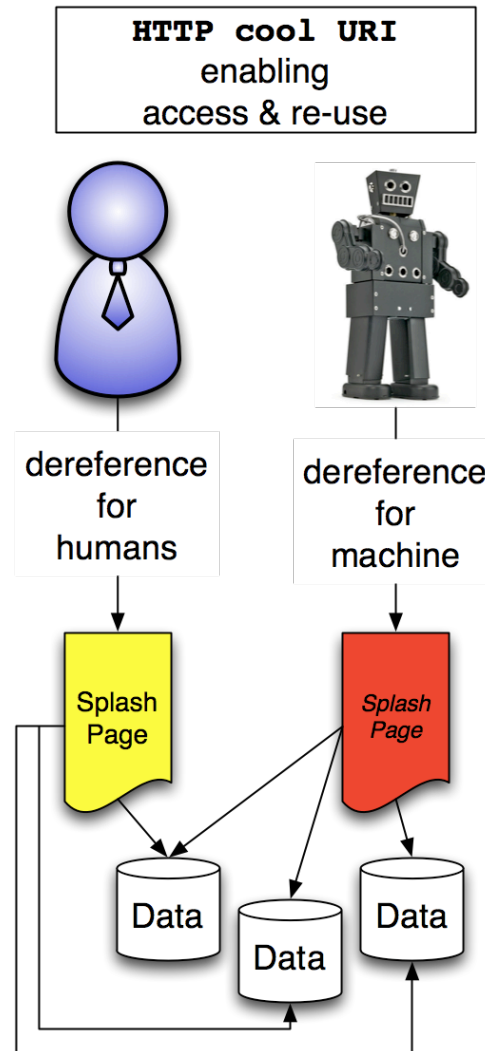
Catering to Human and Machine Agents



Open human actionable *splash page* **Open** machine actionable *splash page*

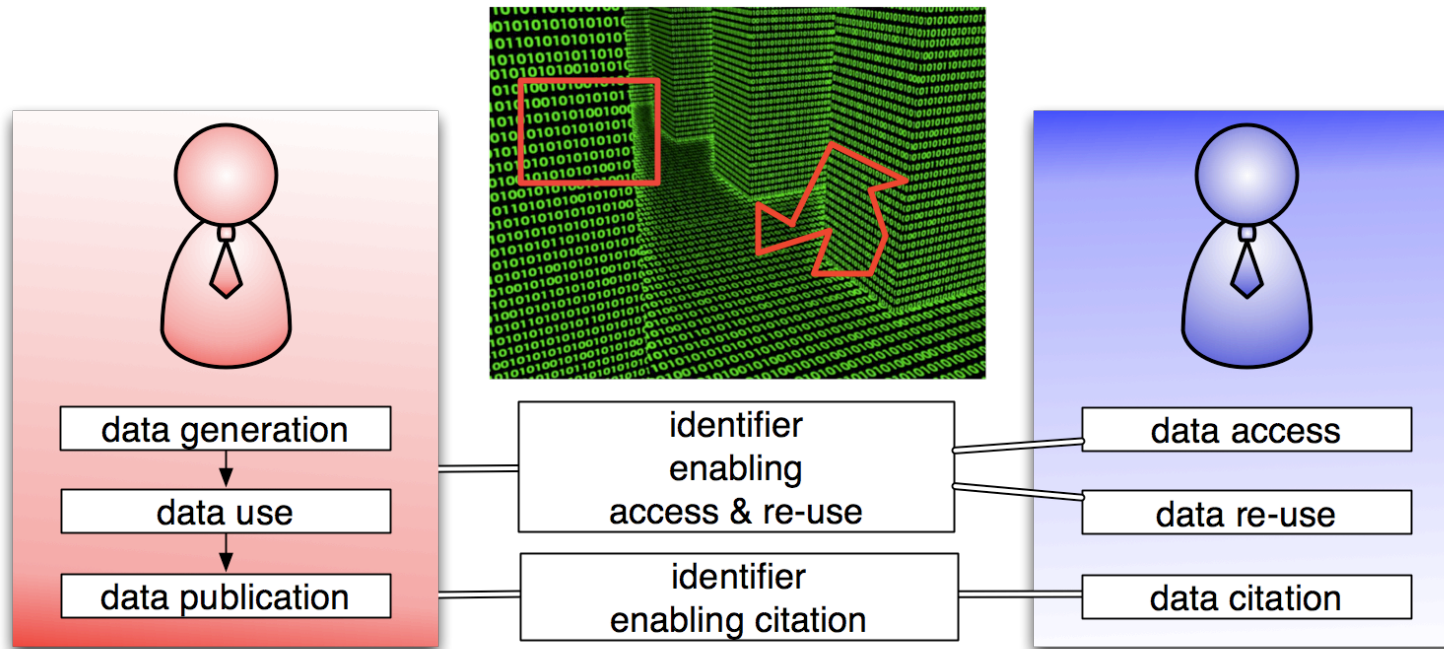


Catering to Human and Machine Agents



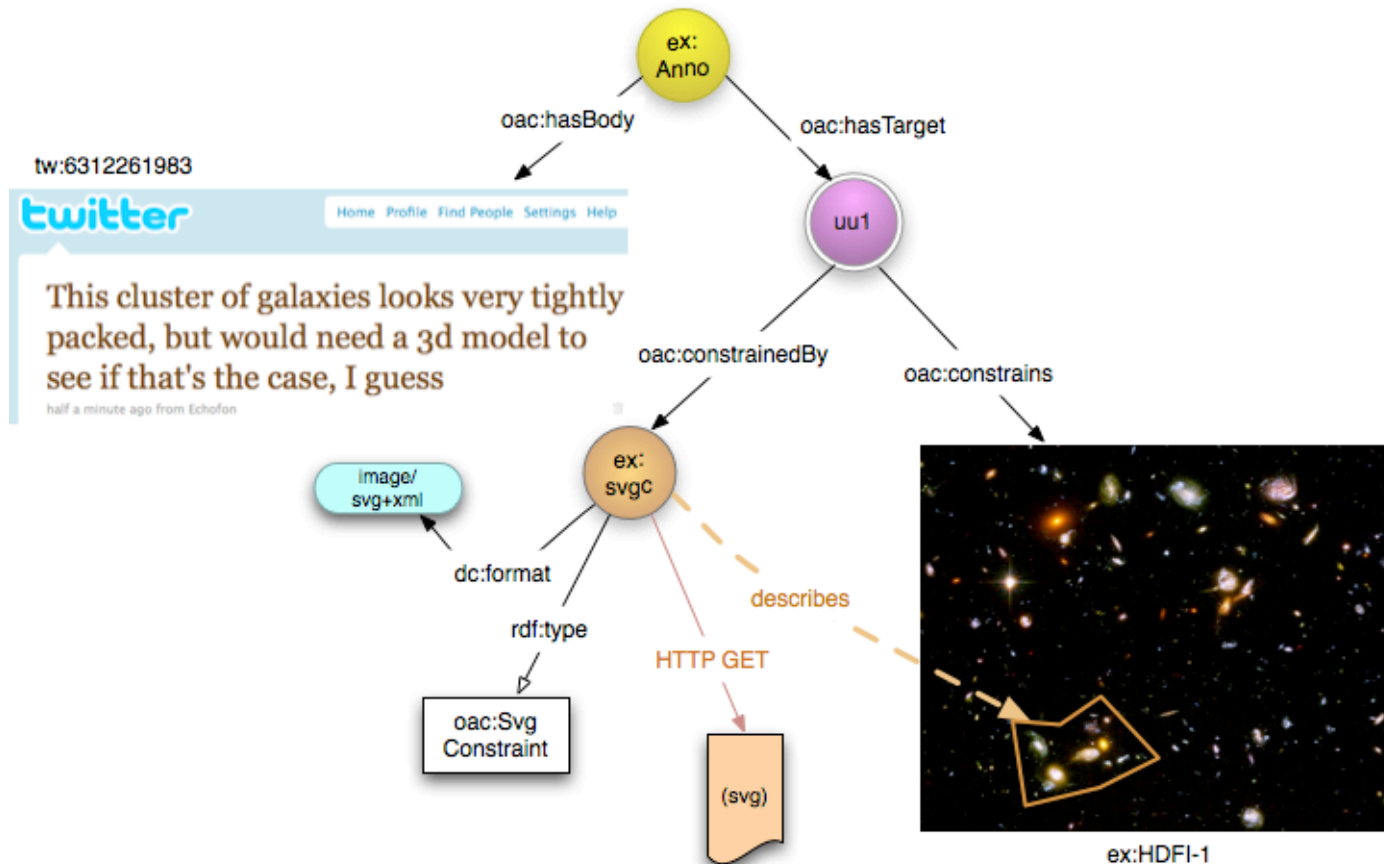
- Attention to machine processing (without human intervention) of data
- Machine actionable metadata, e.g.:
- Clarification of various URIs involved: data (components, formats), splash page, data access points, etc.); cf. OAI-ORE
- Technical metadata about data
- API descriptions
- Authorship (URIs; ORCID)
- Rights & license details
- Provenance account
- ...

Consideration 3: Granularity - Spatial



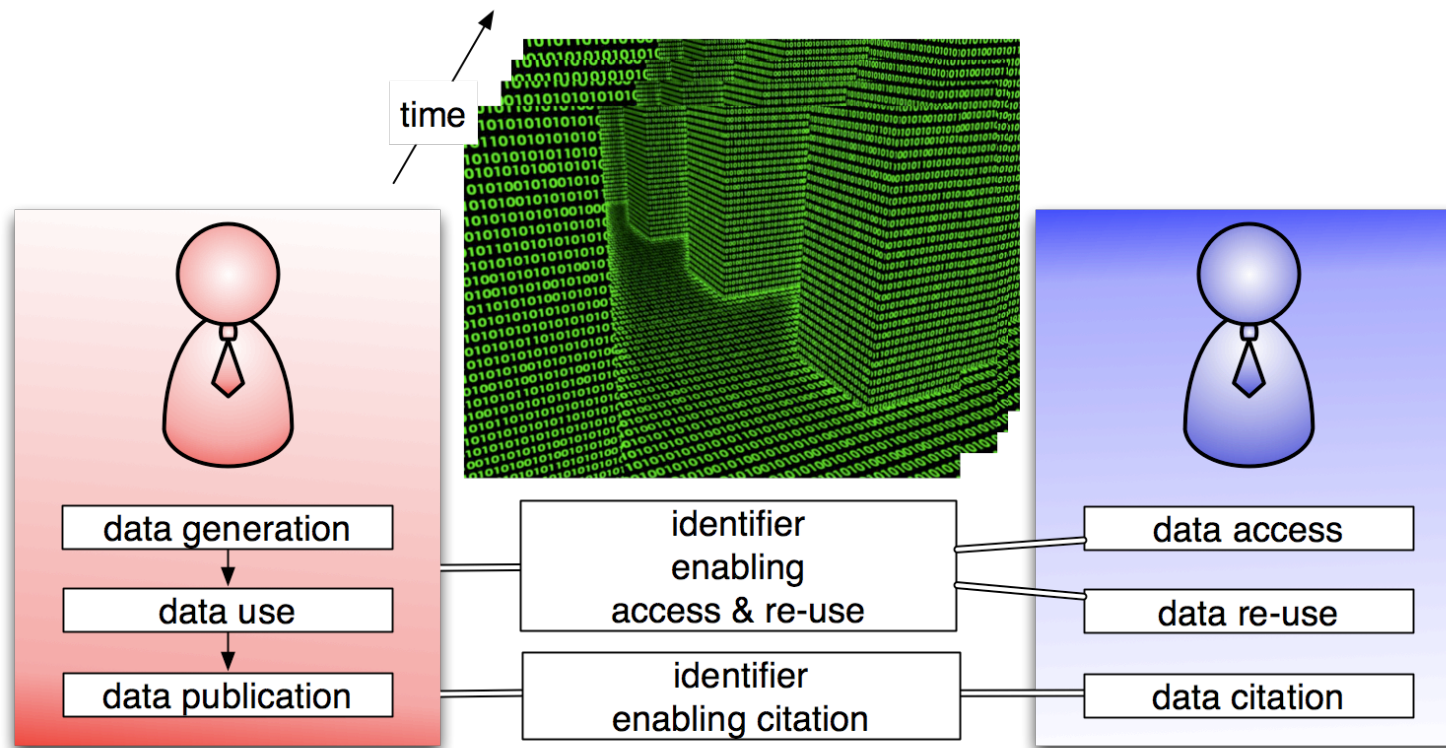
- Simple segments (dimensions in hypercube)
- Complex segments (*arbitrary* queries)

Granularity - Spatial



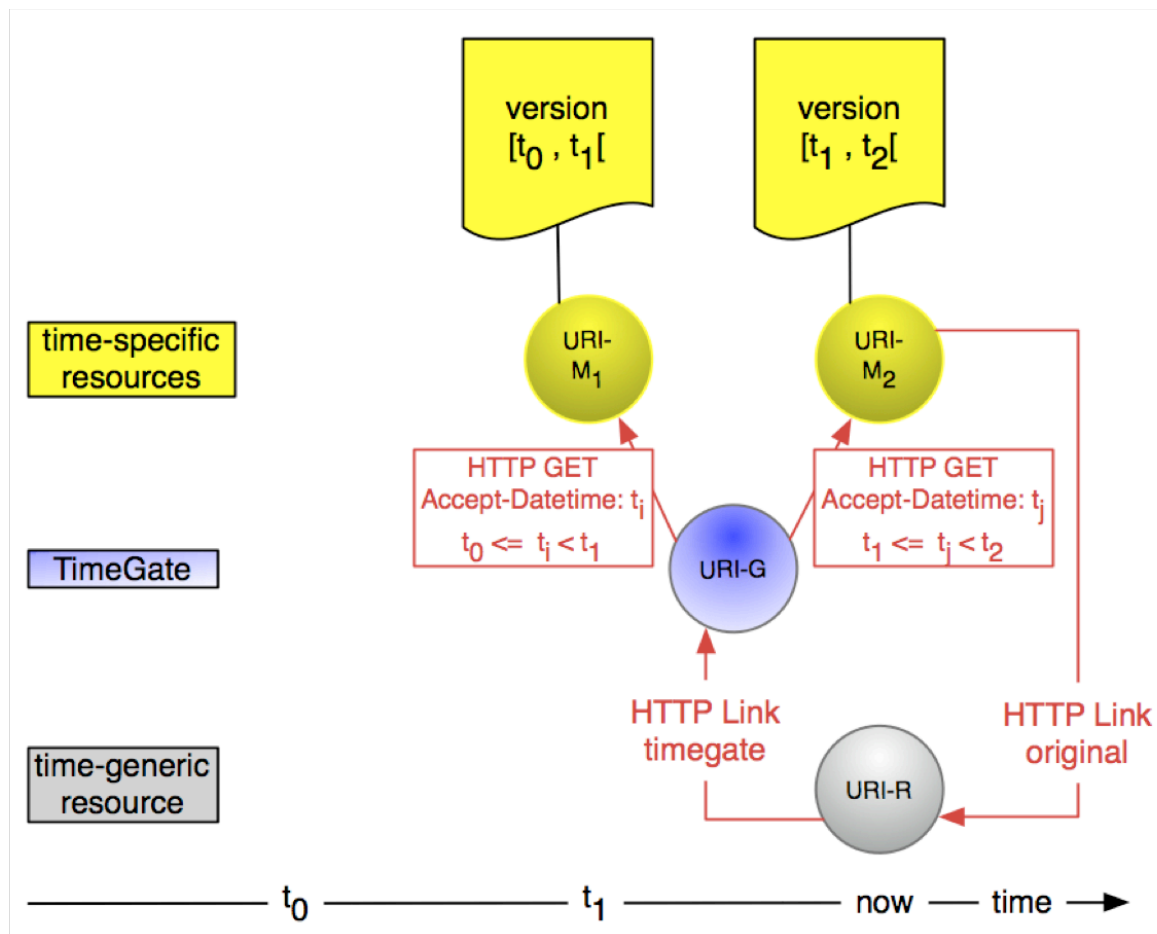
- Similarity with requirement in emerging annotation frameworks to be able to describe segments of resources: Constraints in Open Annotation ; Selectors in Annotation Ontology

Consideration 3: Granularity - Temporal



- Different versions of dataset over time
- Different states of dataset over time (dynamic data)

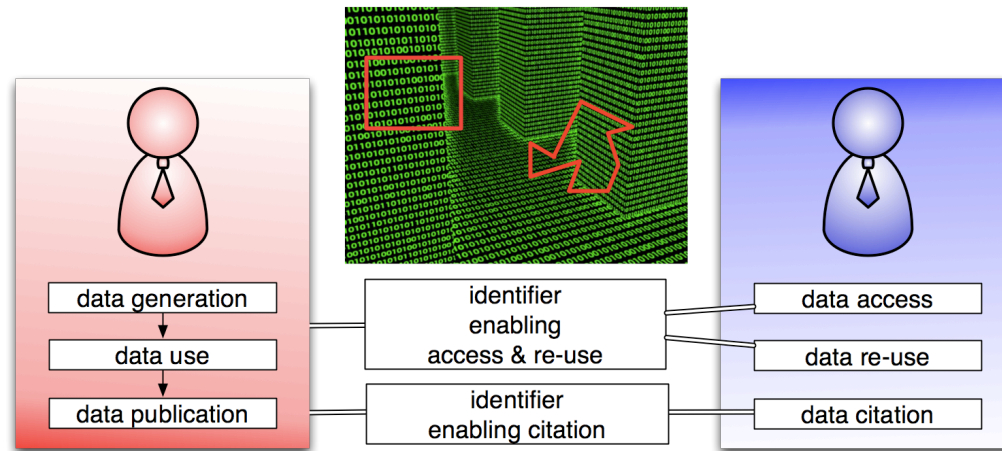
Granularity - Temporal



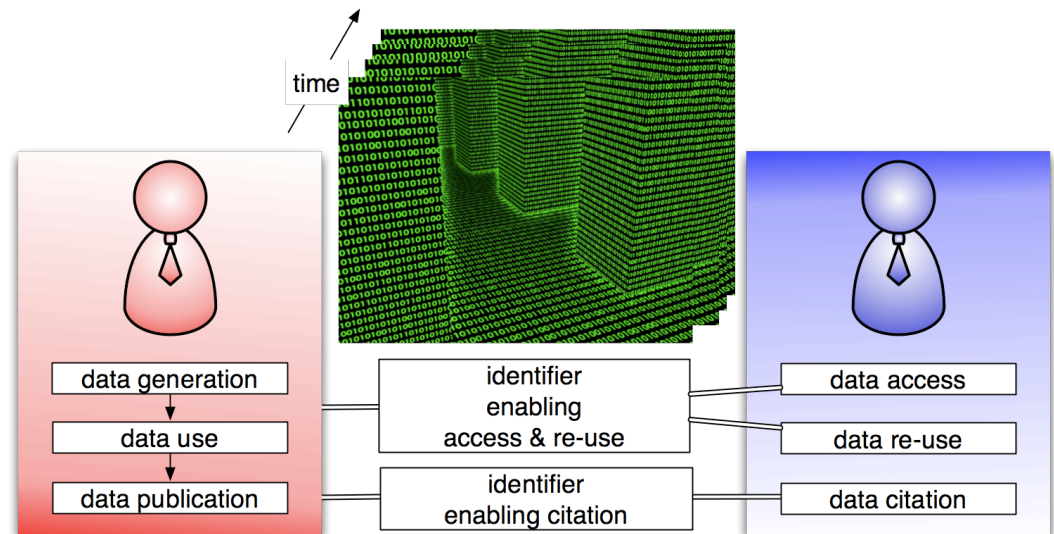
- Similarity with Memento protocol that allows accessing prior versions/states of a resource given its original URI

Granularity – Spatial & Temporal

- Both data publisher and data re-user require granular addressing capability



- There's something to be said in favor of citing at the level of the entire dataset (identifier of data) and adding granular specificity for access and re-use.



Granularity – Spatial & Temporal

- Option:
 - For citation, access, and re-use: Mint new identifier for each segment
- Option:
 - For citation, access, and re-use: Use (entire) dataset identifier with query component
- Option:
 - For citation: Use (entire) dataset identifier
 - For access and re-use: Use additional URI that refers to a constraint/selector on the dataset that describes the segment (both spatial and temporal).
 - Citation becomes a tuple: [dataset URI; constraint URI]



3 Considerations that Merit More Attention

- Nature of Identifiers for Citation, Access, Re-Use
- Catering to Human and Machine Agents
- Granularity: Spatial ; Temporal

