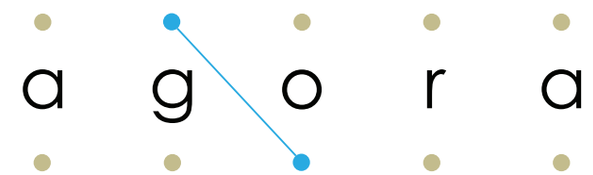
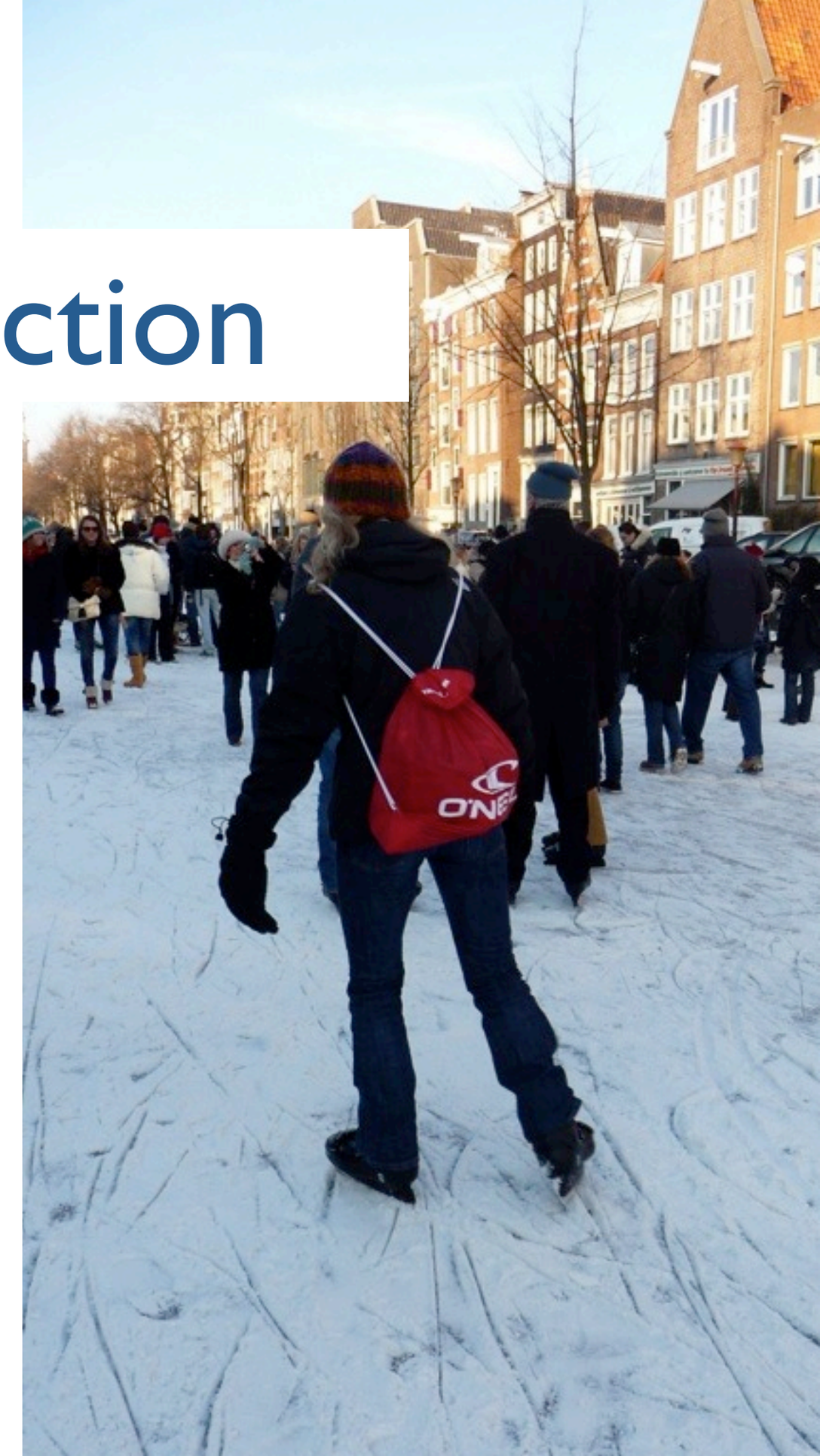
Reusing Linguistic Resources

Tasks and Goals for a Linked Data Approach

Marieke van Erp
marieke@cs.vu.nl



Introduction



- BA, MA & PhD compling/ information extraction @Tilburg University

- Since 2009: SemWeb group @VU University Amsterdam

Why Reuse Linguistic Resources?

- Linguistic resources are expensive to create
- ...and difficult to use for ‘outsiders’

- How can we reach out to the ‘outside world’?

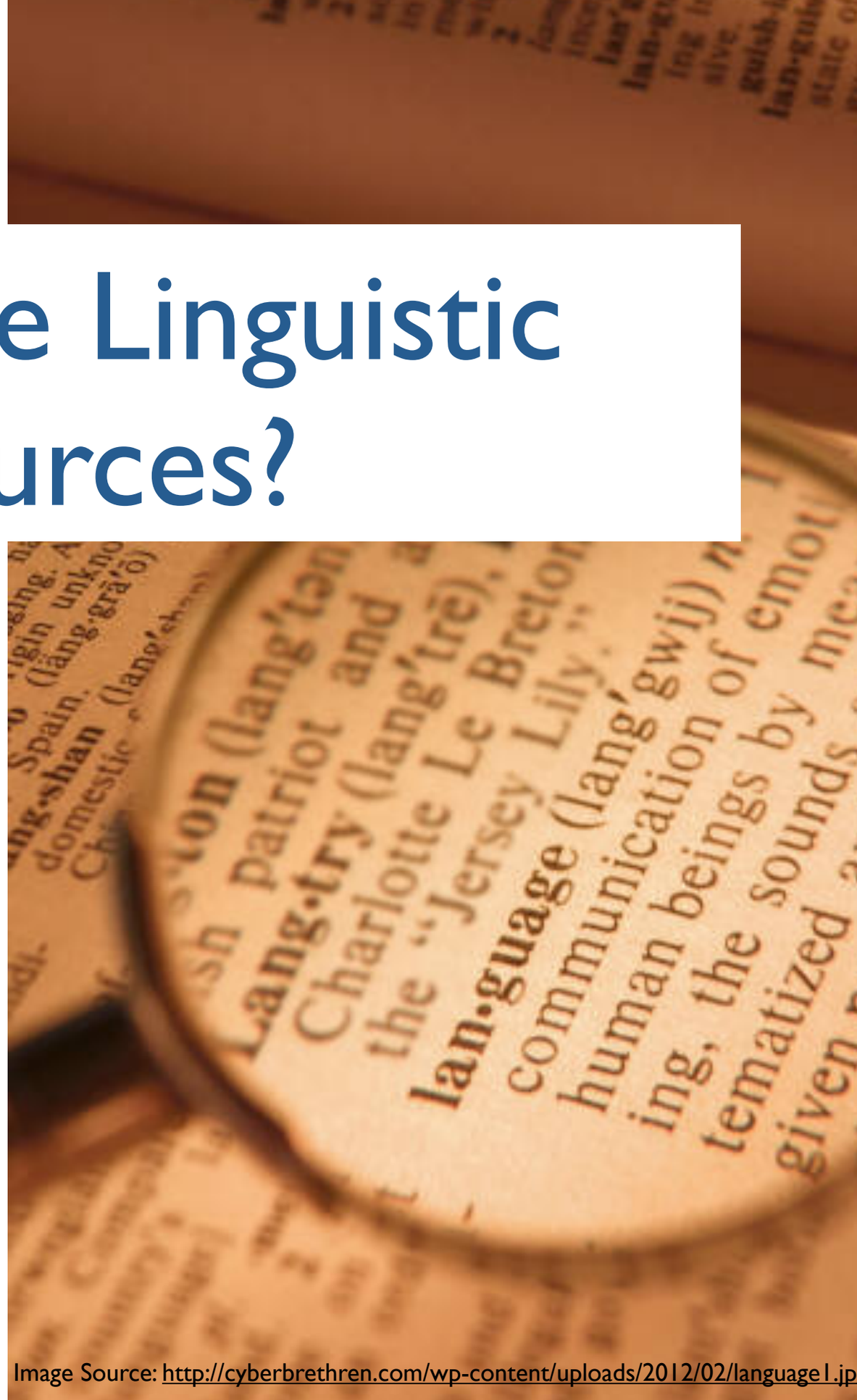


Image Source: http://cyberbrethren.com/wp-content/uploads/2012/02/language1.jp

Make reuse easier!

- Increased visibility
- Social value:
    - stimulates collaboration
    - accelerates innovation
- External quality control

Image Source: http://th02.deviantart.net/fs71/PRE/i/2010/146/b/3/DON__T_PANIC_by_VigilantMeadow.jpg

What's holding us back?

- Fear?
- Habit?

Image Source: h http://mindfulbalance.files.wordpress.com/2011/02/hesitate1.jpg

Practical Constraints

1. Task specificity
2. Formats
3. Different conceptual models
4. No machine-readable definitions
5. Lack of metadata

Image Source: http://bogdankipko.com/wp-content/uploads/2011/12/barriers.jpg

1. Task-specificity

- Resources are often geared towards one specific task
  e.g., part-of-speech tagging, named entity recognition

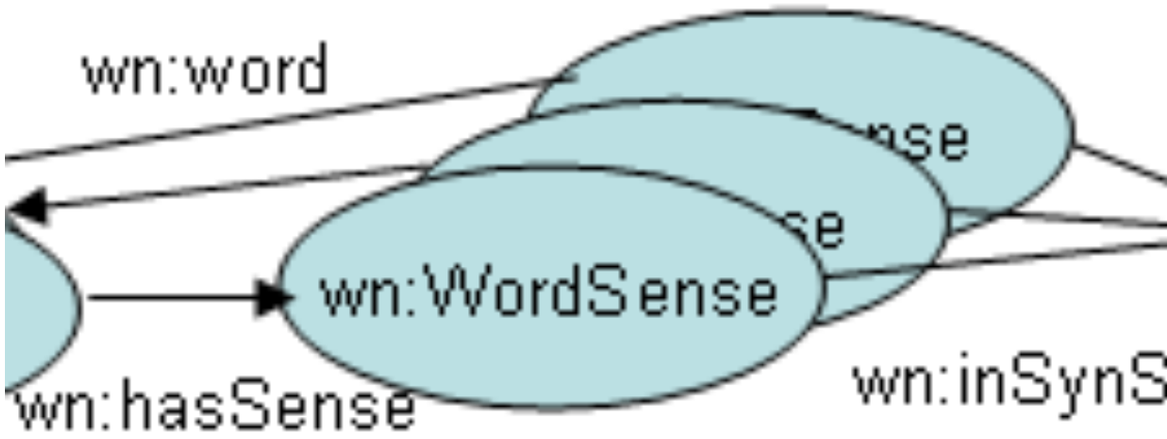
- *How can we make our resources more flexible?*

Image Source: http://thelearnersguild.files.wordpress.com/2008/07/the-informal-learners-toolkit1.jpg

2. Formats

- XML, inline XML, CSV, one word per line, one sentence per line, slashtags, ARFF,

Image Source: http://www.elec-intro.com/EX/05-13-03/kf_compact_data.jpg



3. Conceptual Models

- An NP is an NP is an NP?
- "President Obama signed the National Defense Authorization Act after months of debate"
  - NE: "President Obama"?
  - NE: "Obama"?

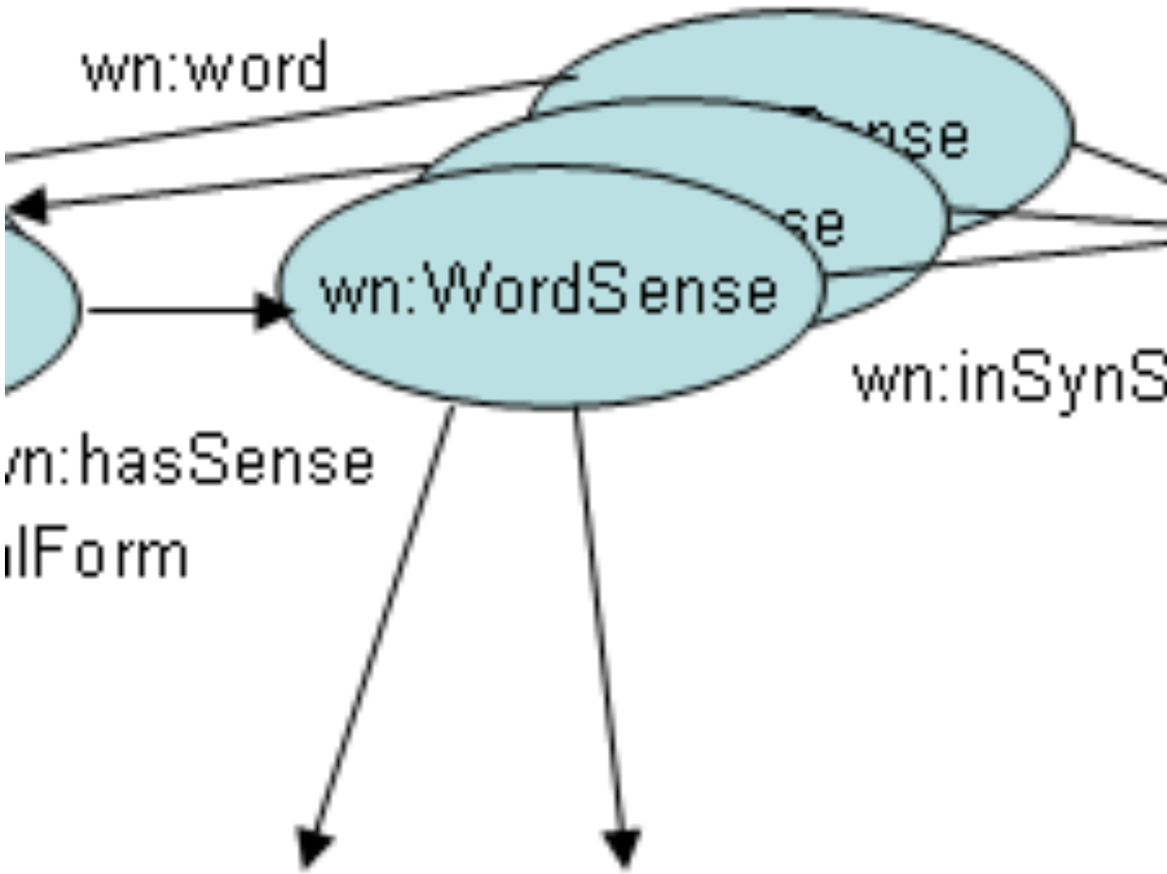


Image Source: http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713-fig01.png

4. Lack of Machine-Readable definitions

- For integration or reuse manual effort is needed
  - time consuming
  - difficult to track definitions
  - not scalable

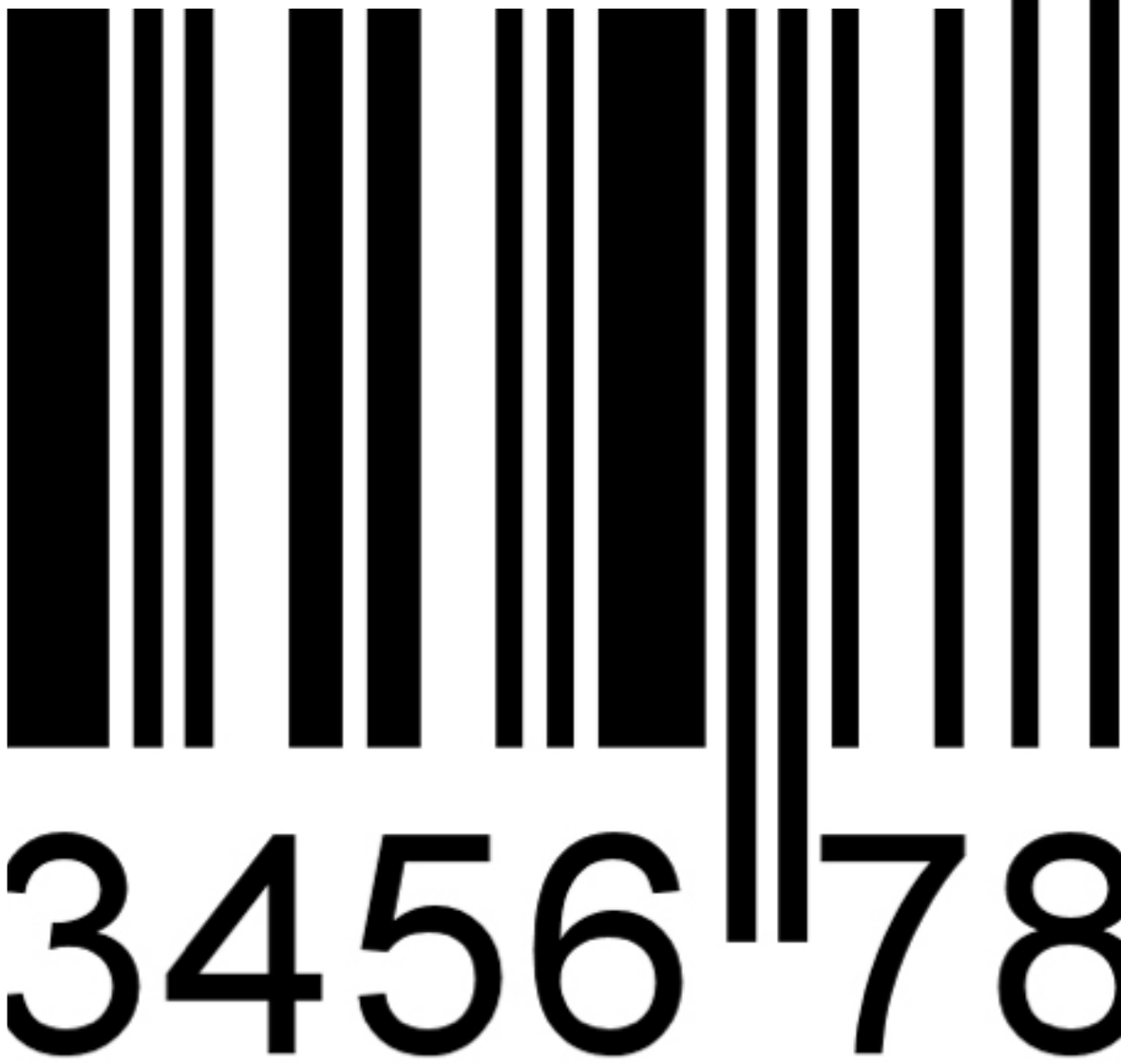


Image Source: http://www.barcode1.co.uk/images/samplejplarge.jpg

5. Lack of Metadata

- Can I trust this data provider?
- How was this data created?
- How many annotators?
  - for the entire data set?
  - per instance?
- If generated automatically, what were the parameters?

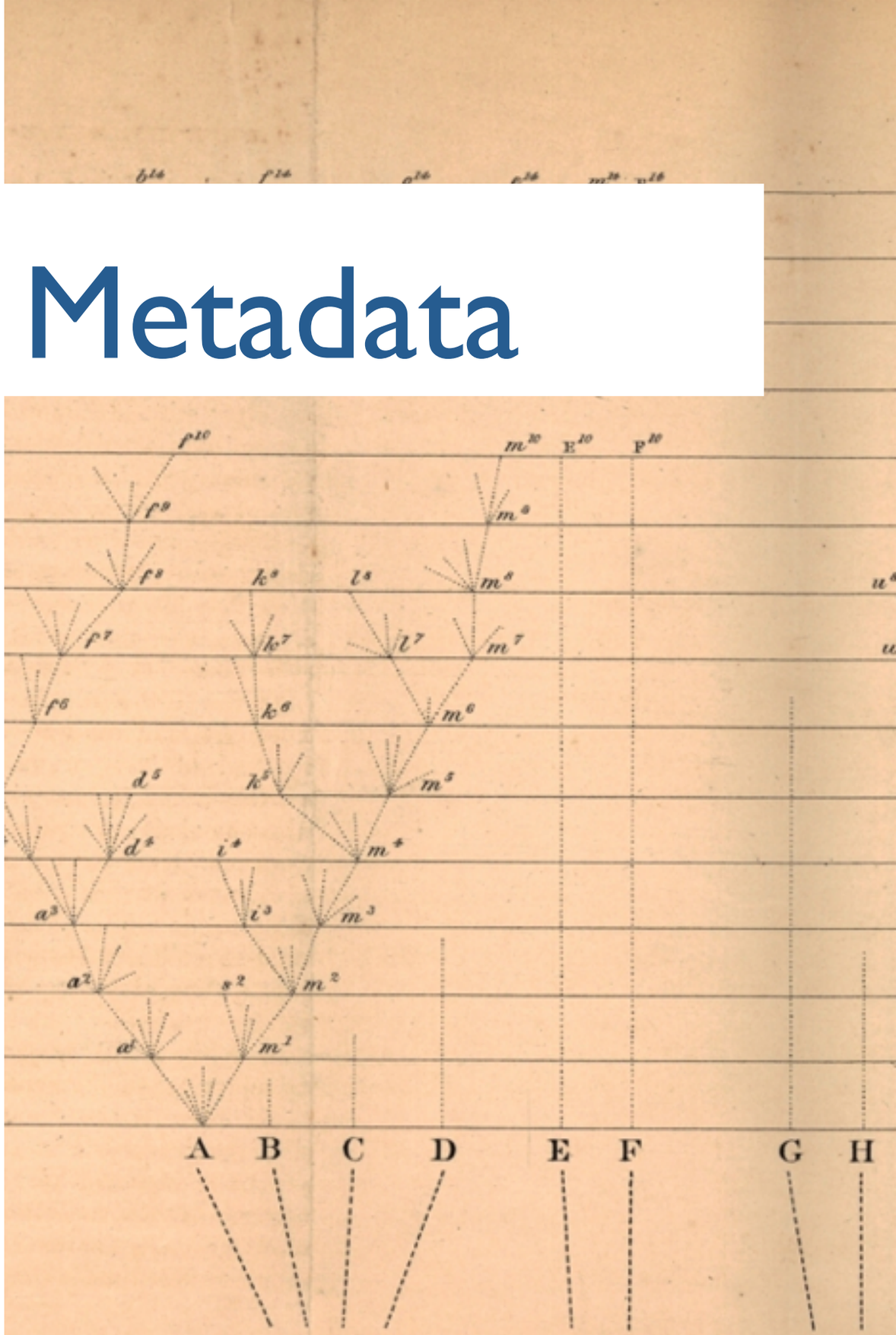


Image Source: http://darwin-online.org.uk/converted/published/1859_Origin_F373/1859_Origin_F373_fig02.jpg

A Linked Data Approach

- Linked Data is not a magic solution to all problems
- ...but it is better than what we've got at this moment

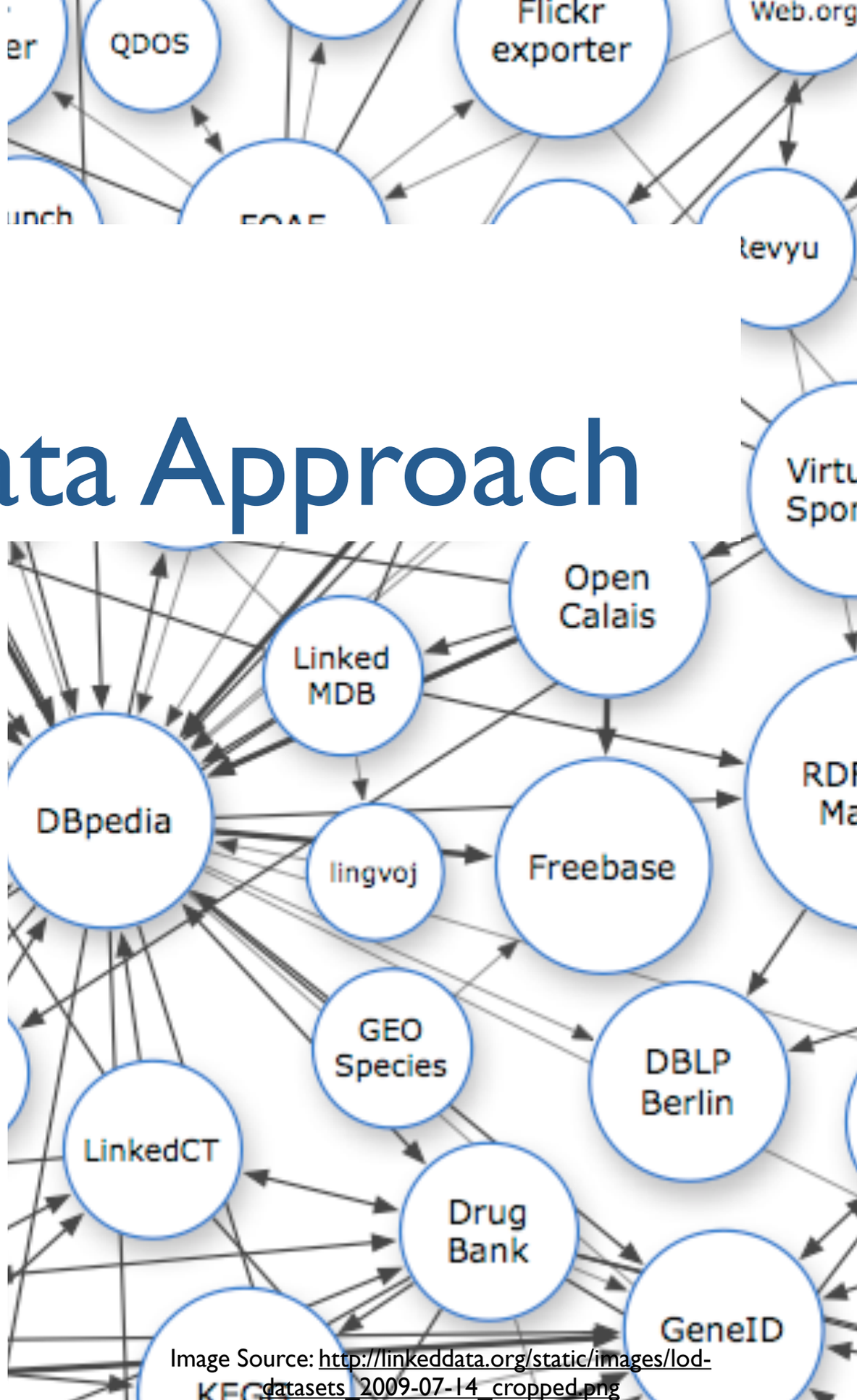


Image Source: http://linkeddata.org/static/images/lod-datasets_2009-07-14_cropped.png

1. Using RDF

- RDF is not inherently better than some other formats, but it is used by many

- + SPARQL makes it easy to retrieve data

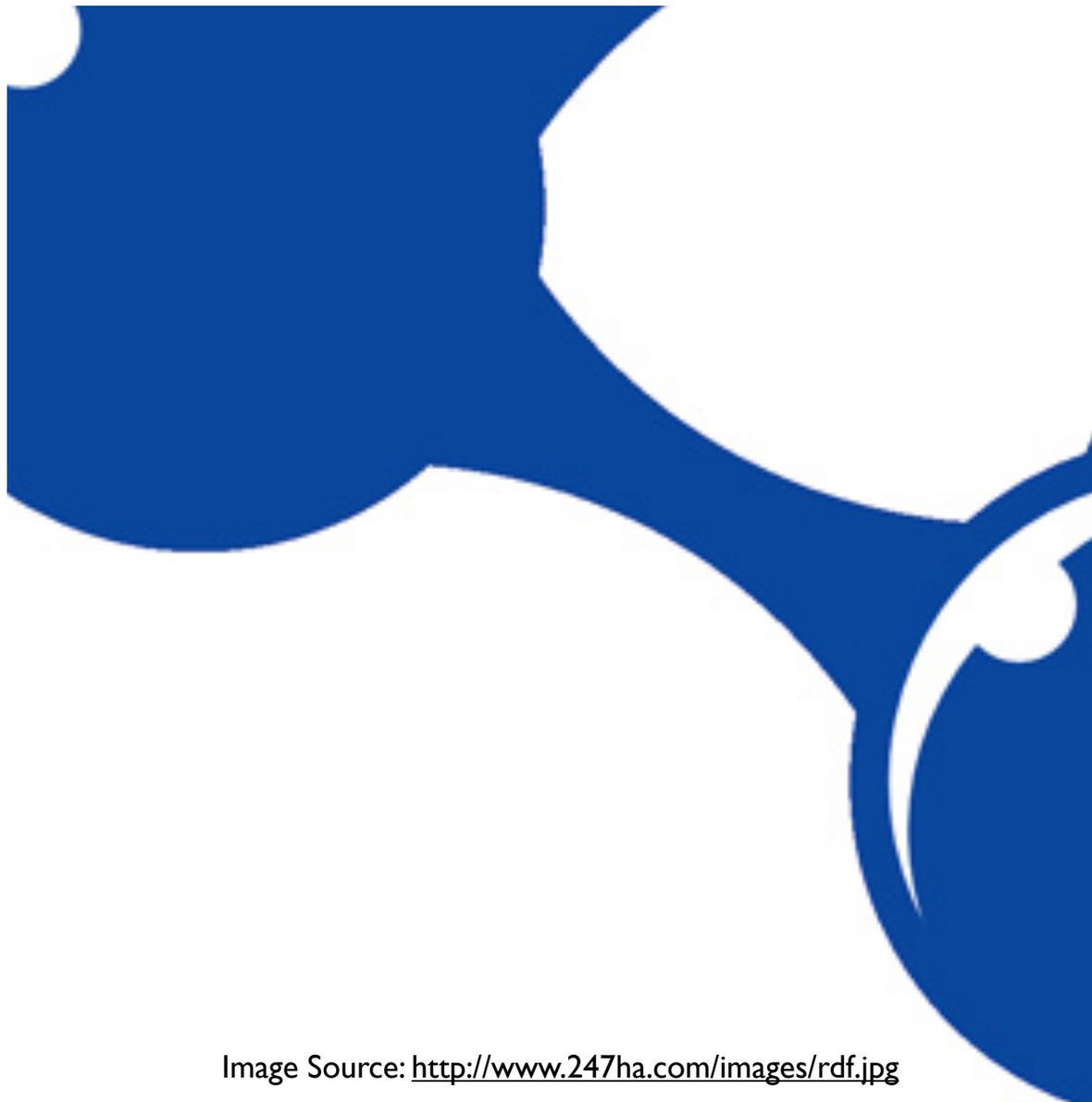


Image Source: http://www.247ha.com/images/rdf.jpg

2. Mapping Annotations

- A single conceptual model for all linguistic resources is not going to happen
- ...but can we spot the similarities between models and utilise that?

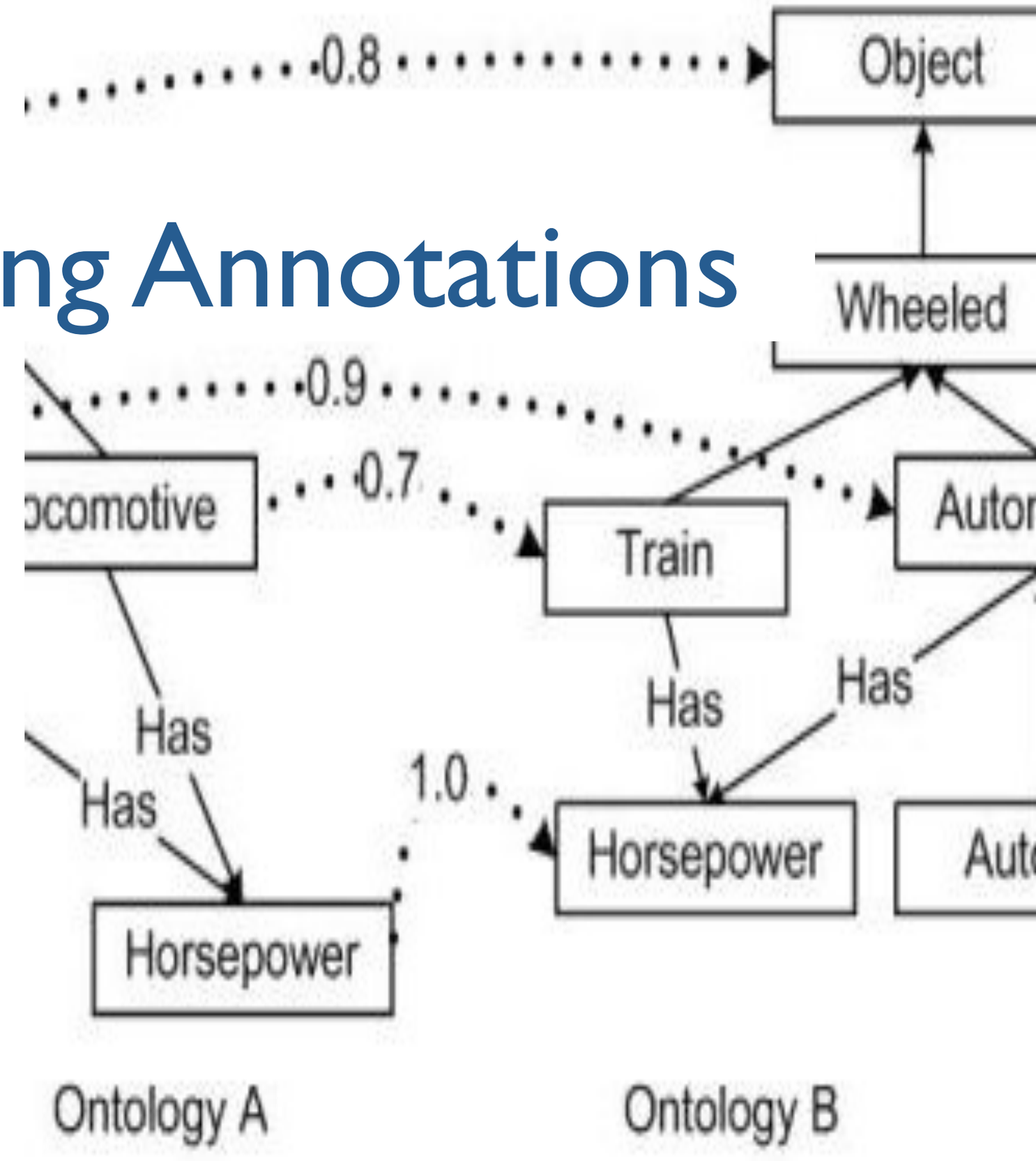


Image Source:http://www.webology.org/2006/v3n3/images/sample.JPG

3. Grounding

- It's only linked data if you link it to other sources

- Added bonus: automatic sense disambiguation + access to a wealth of extra knowledge about your data item

Image Source: http://mj-services.com/wallpaper/More_WallPaper/Trees/Giants,%20Calaveras%20State%20Park%20-%201600x1200%20-%20ID%2015.jpg

4. Define Your Metadata

- Include your data model
- Preferably give each instance's provenance
    - collection
    - annotation/creation
    - previous versions
    - confidence

Image Source: http://www.wineaustralia.com/australia/Portals/2/November%20E-news/Wines%20of%20Provenance%20Final.jpg

Conclusions

- Look for similarities between resources
- Say where your resource comes from
- Use standards, or make it easy for others to convert your data to a standard
- Link to other data

Image Source: http://efr0702.files.wordpress.com/2012/02/puzzle.jpg

Questions?

marieke@cs.vu.nl
http://www.cs.vu.nl/~marieke

Image Source: http://www.amichelleblakeley.com/storage/question%20marks.jpg?__SQUARESPACE_CACHEVERSION=1295297003883

Acknowledgment

- This work is funded by NWO in the CATCH programme, grant 640.004.801