Dynamic Data Citation Enabling Reproducible Research

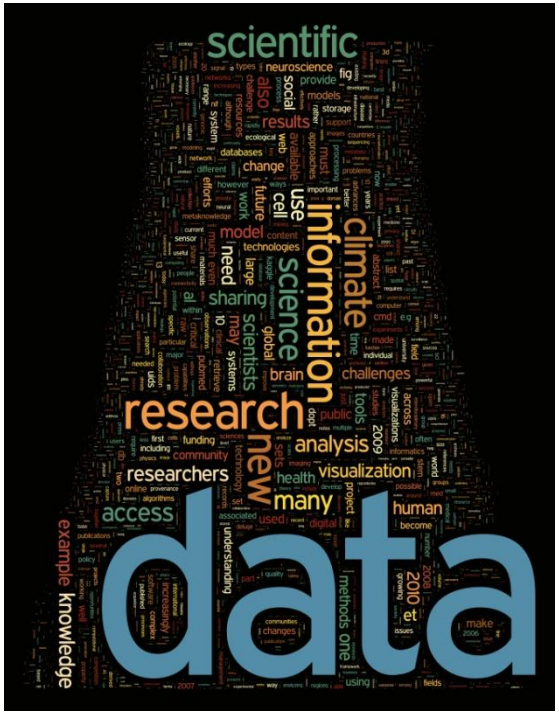
Stefan Pröll, Andreas Rauber

sproell@sba-research.org

Open Space for ICT

27.11.2014

Data and Data Citation



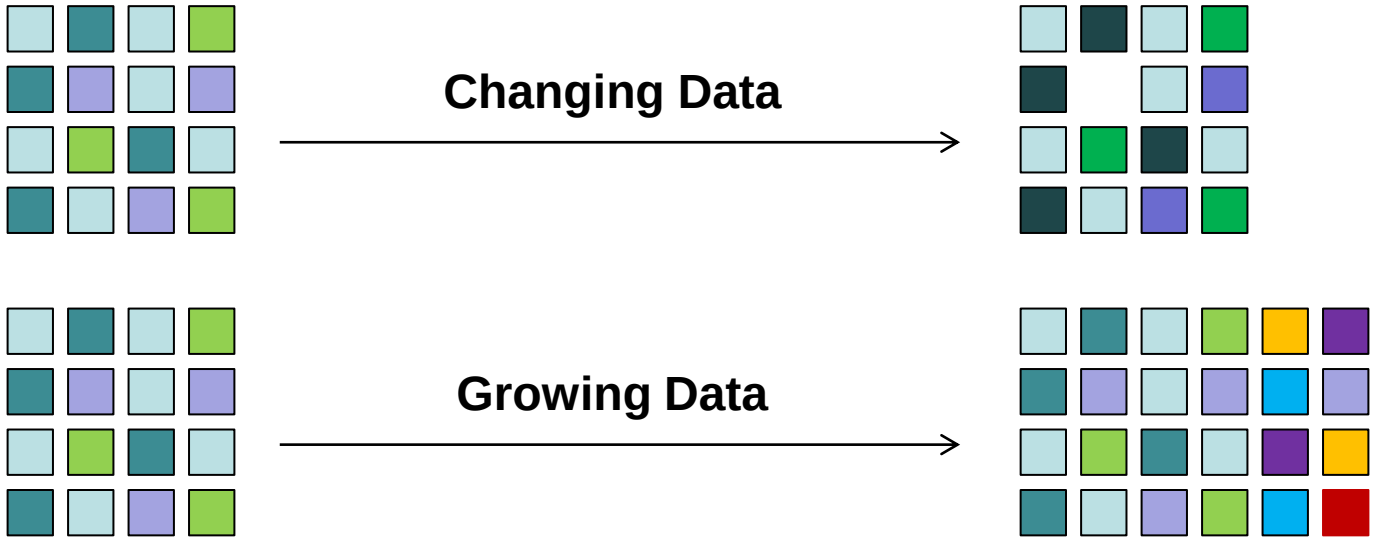
http://www.cccblog.org

- Data is essential for experiments
  - Reproducibility
  - Verification
  - Validation
- Data as a "1st-class citizen" in science
- We need to be able to
  - **preserve** data and keep it accessible
  - cite data to **give credit** and show which data was used
  - **identify precisely** the data used in a study/process for repeatability, verifyability,…

How are Datasets Cited Today?

- Persistent Identifier (PID) e.g. DOI, URI, ARK, … currently provided for
  - entire data sets, copies of subsets
  - static data, sometimes releases of versions (annual etc)
  - cited in their entirety with textual description of subsets
- This is insufficient in many settings
  - not machine-actionable
  - not scalable for large data sets
  - insufficient support for data that changes
  - insufficient support for arbitrary subsets (rows/columns)

Dynamic Data

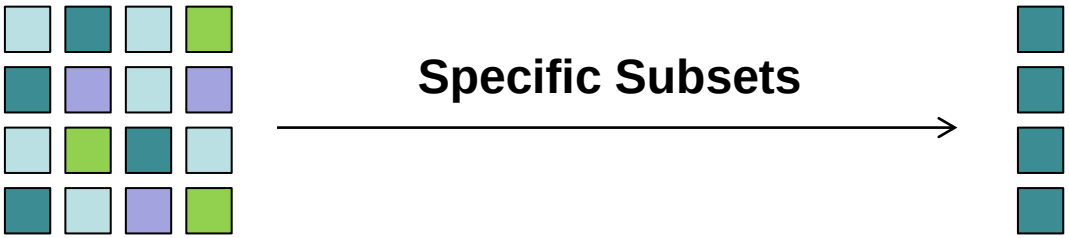


Research data is **not stable**. Experiments require adaptations of algorithms, software, parameters, settings …

- New data is produced
- Existing data gets updated

Dynamic data needs to be citable and accessible

Specific Subsets



Researchers require **specific subsets** of data

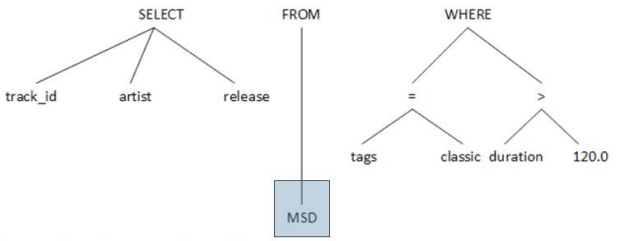
- Selected data required in different granularities
- Creating subsets requires domain knowledge
- Storing individual data exports for each subset and version is not feasible

Retrieving the very same subset from dynamic data is not trivial.
We need:

- Versioning and timestamping of data
- Query mechanisms

Making Data Citable

- What is needed:
  - Time stamps of data
  - Versioned data:
    - Record events
    - No actual deletes
      - Unless required by law
  - Query centric data citation
    - Precise query language for constructing subsets
  - Persistent query store that keeps queries and the timestamp of their issuing
  - An identification mechanism for queries, that enables access
    - Assign PIDs to queries




```
SELECT results.track_id, results.artist, results.release
        FROM MSD AS results JOIN (
                SELECT track_id, max(timestamp) AS latestTimestamp
                FROM MSD
                WHERE timestamp <= (SELECT @queryExecutionTimestamp)
                AND (track_id NOT IN
                        (SELECT track_id FROM MSD AS deletedRecords
                                WHERE deletedRecords.status_mark = 'deleted'
                                AND (deletedRecords.timestamp < @queryExecutionTimestamp))
                        )
        GROUP BY track_id
) AS version ON results.track_id = version.track_id AND results.timestamp = version.latestTimestamp

WHERE
    results.tags = 'classic'  AND results.duration> 120
ORDER BY results.track_id;
```

Pilot Studies, Communities and Initiatives

- Natural Environment Research Council (NERC)
  - Environmental Change Network
  - Long-term environmental monitoring from automatic and manual recording across the UK
- Additional pilots: http://rd-alliance.org/groups/data-citation-wg/wiki/collaboration-environments.html
- Planned workshops to evaluate our solutions:
  - Earth Science Information Partners (ESIP)
    - Workshop in Washington Jan 8 2015
  - European Space Agency
    - Workshop April 2015

Questions?

- My questions: How can we apply data citation in the domain of space data? How to improve reproducibilty of scientific experiments? How can be share data and enable easy access?

Thank you for your attention!