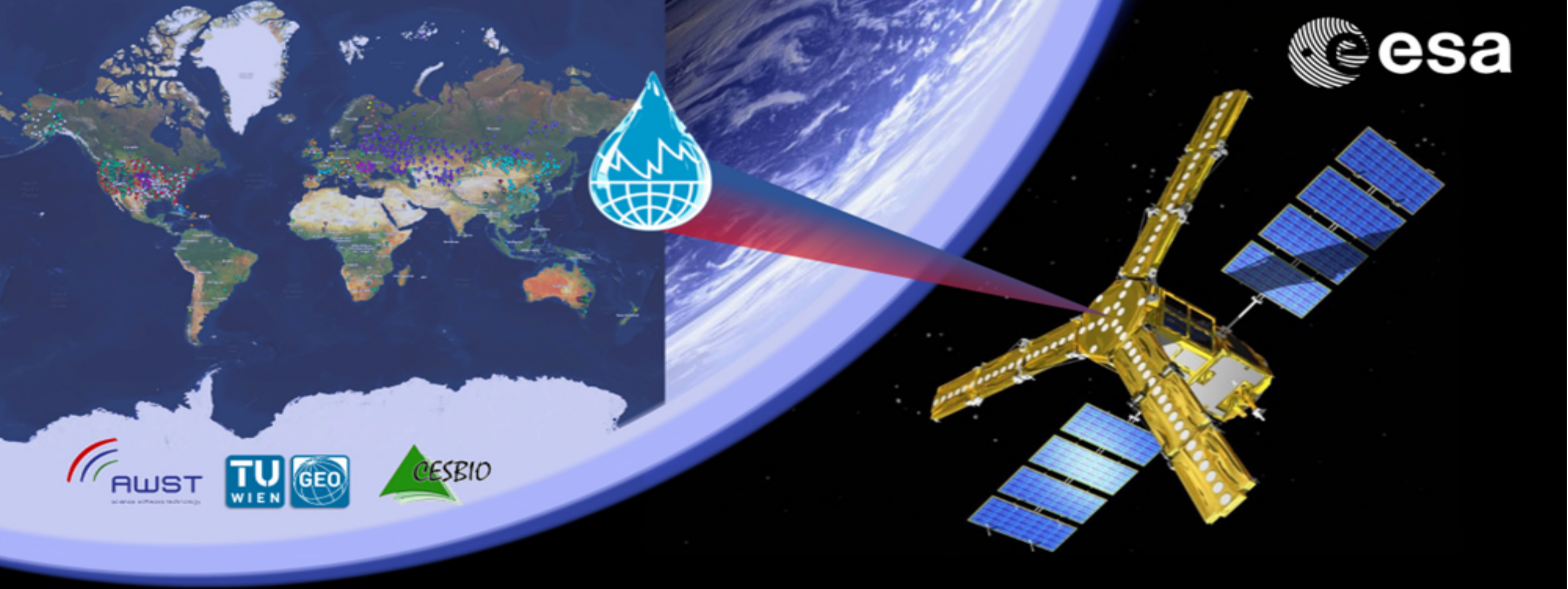


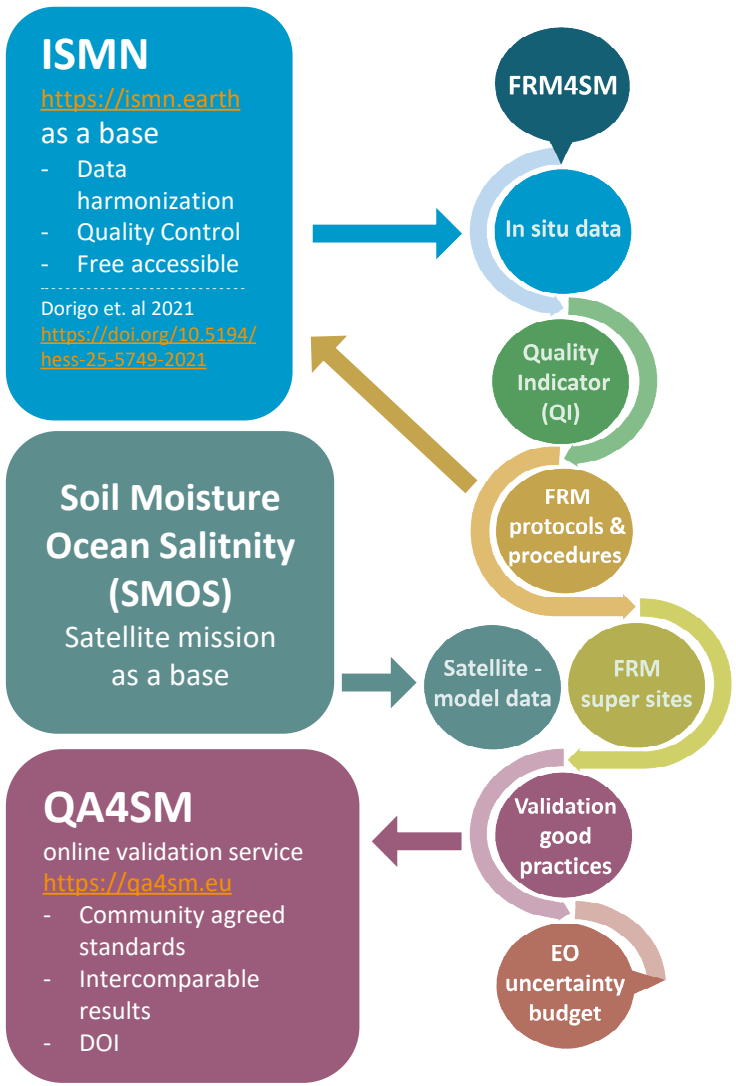
FRM4SM Data traceability

Fiducial Reference Measurement for Soil Moisture II May 2021 – May 2023

Irene Himmelbauer[1], **Moritz Staudinger[5], Tobias Hasjan[5]**, Andreas Rauber[5], Tomasz Miksa[5], Daniel Aberer[1], Alexander Gruber[1], Wolfgang Preimesberger[1], Pietro Stradiotti[1], Wouter Dorigo[1], Monika Tercjak[2], Alexander Boresch[2], Arnaud Mialon[3], Francois Gibon[3], Philippe Richeaume[3] , Yann Kerr[3], Raul Diez Garcia[4], Raffaele Crapolicchio[4], Roberto Sabia[4], Klaus Skipal[4], Philippe Goryl[4]

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Short introduction to FRM4SM = Fiducial Reference Measurement for Soil Moisture



- 4 partners ([1]TU Wien, [2]AWST, [3]CESBIO, [4]ESA)
- Scientific Advisory group: 10 members
- Building upon community agreed standards

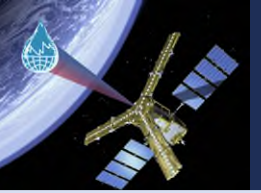
→ More info on https://project-frm4sm.geo.tuwien.ac.at/

EGU session HS6.1 (all Friday)

Data Traceability:

- Uncertainty understanding
- Data → FAIR principles → tracking data changes

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway



ISMN status quo – issues to go towards FAIR

https://ismn.earth

- Not versioned
- Dynamically changing data
  - Updates: nrt, regular, irregular, historical datasets
  - Reprocessing of data (overwriting current data)
- No Persistent Identification number (PID) attached (data/download)
- Licence still “copyright“ – no onward distribution allowed
  - ISMN data provider community agreed
  - Outsourcing the data for PID issuing not possible (e.g., Zenodo, etc.)
    - ISMN data usage statistics important – ½ yearly provider reports
    - Could mandate be lost when outsourcing the data?
- ISMN data accessible when registered/logged in (all for free)
  - Security --- locations are not allowed to be forwarded (stolen material)
  - ISMN data provider community agreed

Can the ISMN go towards FAIR?

Answer: YES

Department for Informatic, Information and Software Engineering (TUW)

- **Professor Andreas Rauber & Dr. Tomasz Miksa**
- Co chair of: Working Group on Data Citation (WGDC)

RDA RESEARCH DATA ALLIANCE

References ISMN system currently built upon:

[1] FAIR: https://www.go-fair.org/fair-principles/
[2] WGDC: https://zenodo.org/record/1406002/files/datacitation.pdf?download=1
[3] Rauber et al. 2021: https://doi.org/10.1162/99608f92.be565013
[4] Data Cite metadata Schema 4.4 https://doi.org/10.14454/3w3z-sa82

Going FAIR / making data traceable without not wanting to OR not being able to outsource the data

→ "Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data" [3]
  → In house versioning - full reproducibility / track changes
  → Versioning system built upon recommendations of WGDC [2]
→ In house DOI system for ISMN downloads [4]

ISMN evaluated and tested for:

→ In house versioning of data changes [3]:
  → full reproducibility of soil moisture data

Master thesis: **Moritz Staudinger**
"Reproducible Querying in evolving, schema changing databases to enhance FAIRness of research data"

→ In house DOI system of data download requests

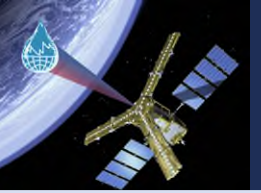
Integrative Project: **Tobias Hajszan** (master`s degree)

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

ISMN in house versioning system

Me?

- BSc in Software and Information Engineering
- Working in Reproducible Research since 2019
  - FAIRness
  - Dynamic Data Citation
  - Reproducible IR-Systems
- Currently MSc in Data Science
- Master Thesis in Cooperation with ISMN



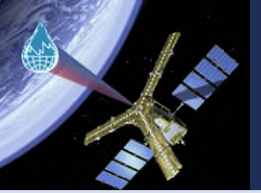
Motivation

- How to cite research data?
  → Get a Persistent Identifier for your dataset

- But what if the data evolves?
  → Create a new dump for every change?
  → Cite the database without mentioning the query used?
  → Not cite the data at all or use an outdated link?

- And now imagine doing this for each possible dataset you are using in your publication!

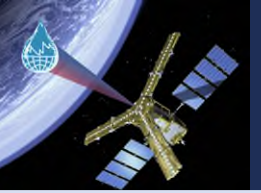
[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway



Why is ISMN interesting?

- 4000 active researchers
- Above 1.4 billion rows
- Constantly evolving database
- Real-Life Big Data example
- Ground Truth for satellite images

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway



Challenges in Data Management

- Data changes without tracking how it was changed
- Downloaded data is not persistently identifiable
- Used subsets are not findable and accessible (FAIR)
- Full database dumps are ineffective in terms of storage space
- The data is not allowed to be onward distributed after download

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Dynamic Data Citation

- Data Versioning
- Query Storing
- Schema Changes
- Data Citation

→ 14 general guidelines published by the RDA in 2016 [1]

→ Meta-analysis about implementations in 2021 [2]

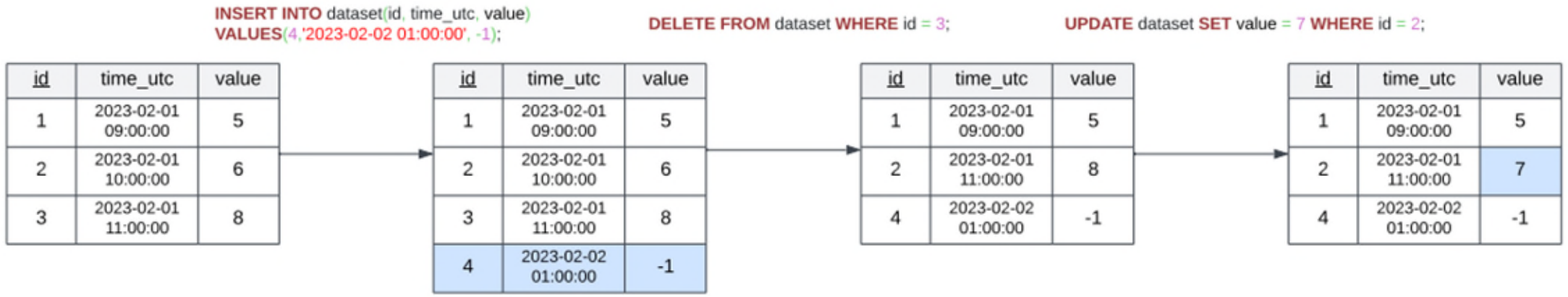
Data Versioning Principles

Three different approaches for tuple based versioning:

- Integrated
- Separated
- Hybrid

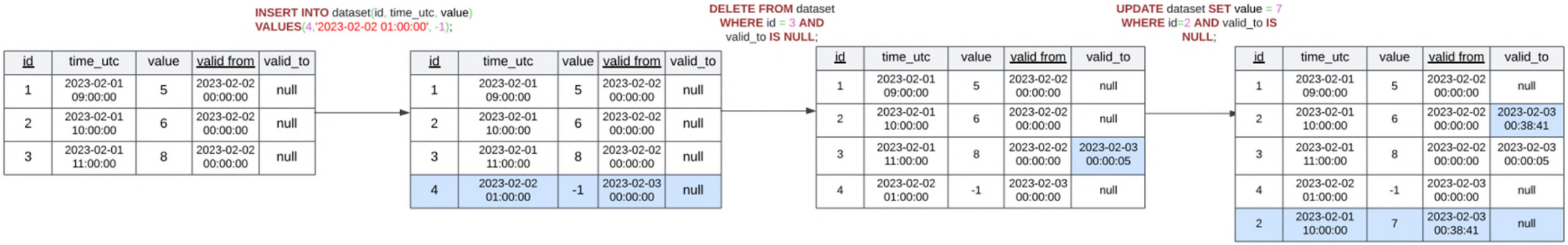
- Add a validitiy period for each tuple
  - On Insert start validity
  - On Delete stop validity
  - On Update perform Delete & Insert

Unversioned Example



[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Integrated Versioning Example



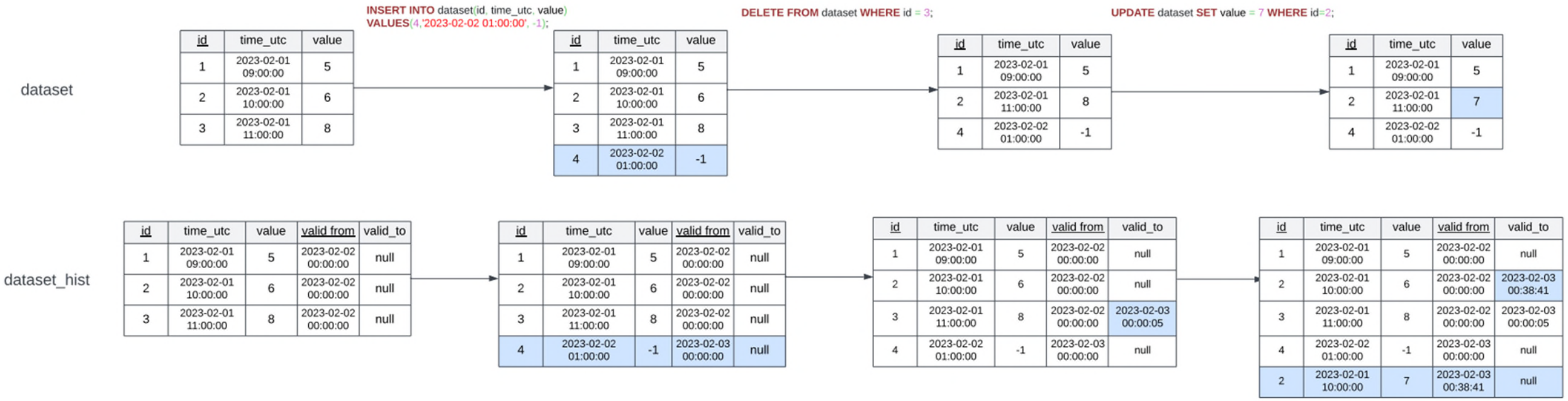
id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	null

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	null
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	2023-02-03 00:00:05
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	2023-02-03 00:38:41
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	2023-02-03 00:00:05
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null
2	2023-02-01 10:00:00	7	2023-02-03 00:38:41	null

Separated Versioning Example



id	time_utc	value
1	2023-02-01 09:00:00	5
2	2023-02-01 10:00:00	6
3	2023-02-01 11:00:00	8

id	time_utc	value
1	2023-02-01 09:00:00	5
2	2023-02-01 10:00:00	6
3	2023-02-01 11:00:00	8
4	2023-02-02 01:00:00	-1

id	time_utc	value
1	2023-02-01 09:00:00	5
2	2023-02-01 11:00:00	8
4	2023-02-02 01:00:00	-1

id	time_utc	value
1	2023-02-01 09:00:00	5
2	2023-02-01 11:00:00	7
4	2023-02-02 01:00:00	-1

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	null

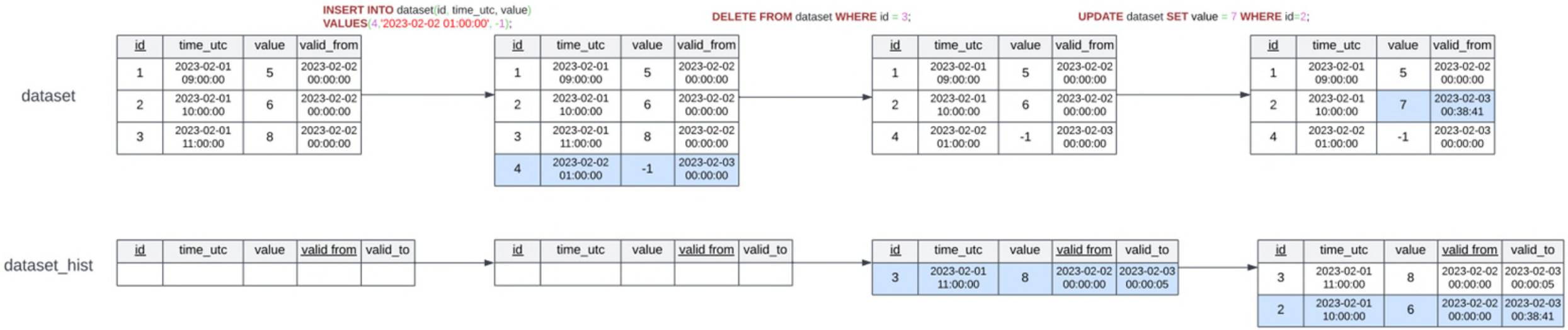
id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	null
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	null
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	2023-02-03 00:00:05
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-02 00:00:00	2023-02-03 00:38:41
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	2023-02-03 00:00:05
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null
2	2023-02-01 10:00:00	7	2023-02-03 00:38:41	null

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Hybrid Versioning Example



Data Versioning Implementation

- In SQL, so no adaptions to workflow necessary
- Overwritten CRUD operations with triggers and rules
- Necessary adaptions
  - Avoid unnecessary updates
  - Rewrite queries to use versioning

Query Store Principles

- Save Execution Time
- Save Query with results number and result hash
- Normalize/Rewrite Query
- On Execution add Timestamp & lock database accordingly

Query Store Principles

id	time_utc	value	valid from	valid_to
1	2023-02-01 09:00:00	5	2023-02-02 00:00:00	null
2	2023-02-01 10:00:00	7	2023-02-02 00:00:00	2023-02-03 00:38:41
3	2023-02-01 11:00:00	8	2023-02-02 00:00:00	2023-02-03 00:00:05
4	2023-02-02 01:00:00	-1	2023-02-03 00:00:00	null
2	2023-02-01 10:00:00	6	2023-02-03 00:38:41	null

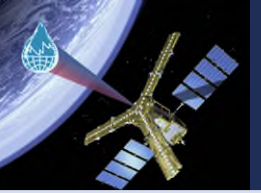
```
SELECT value as 'value',
time_utc as 'time_utc'
FROM dataset;
```

```
SELECT value as 'value', time_utc
as 'time_utc'
FROM dataset
WHERE valid_from <= '2023-02-03
00:40:00' AND (valid_to > '2023-
02-03 00:40:00' OR valid_to IS
NULL);
```

Results: 3
Resulthash: sha256:..54f70
Execution Time: '2023-02-03 00:40:00' …

Query Store Implementation

- Track all user-triggered downloads & save all executed queries
- Execute all queries with timestamps
- Allow to mint a DOI for a specific download
- On re-execution verify result correctness

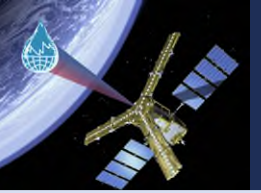


Data Citation Principles

- Combines previous parts
- Allows to Cite/Recreate Subsets of the database
- Persistent Global Identifier for each Query
- Generate Automatic Citations (e.g. Bibtex)

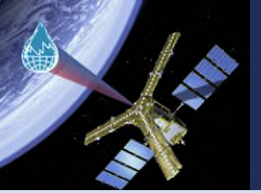
Data Citation Implementation

- Mint DOI per Download, not per Query (as not feasible)
- Landing Page showing all Metadata
- Allow Re-execution of queries via Landing Page
- Automatically give credit to all Networks and Funders in citations



Evaluation Setup

- Iterative Integration of new data of 10 selected networks of various size
- Execution & re-execution of queries after each update
- Measurement of storage increase in absolute and relative size
- Measurement of query-execution, re-execution and database operation speeds

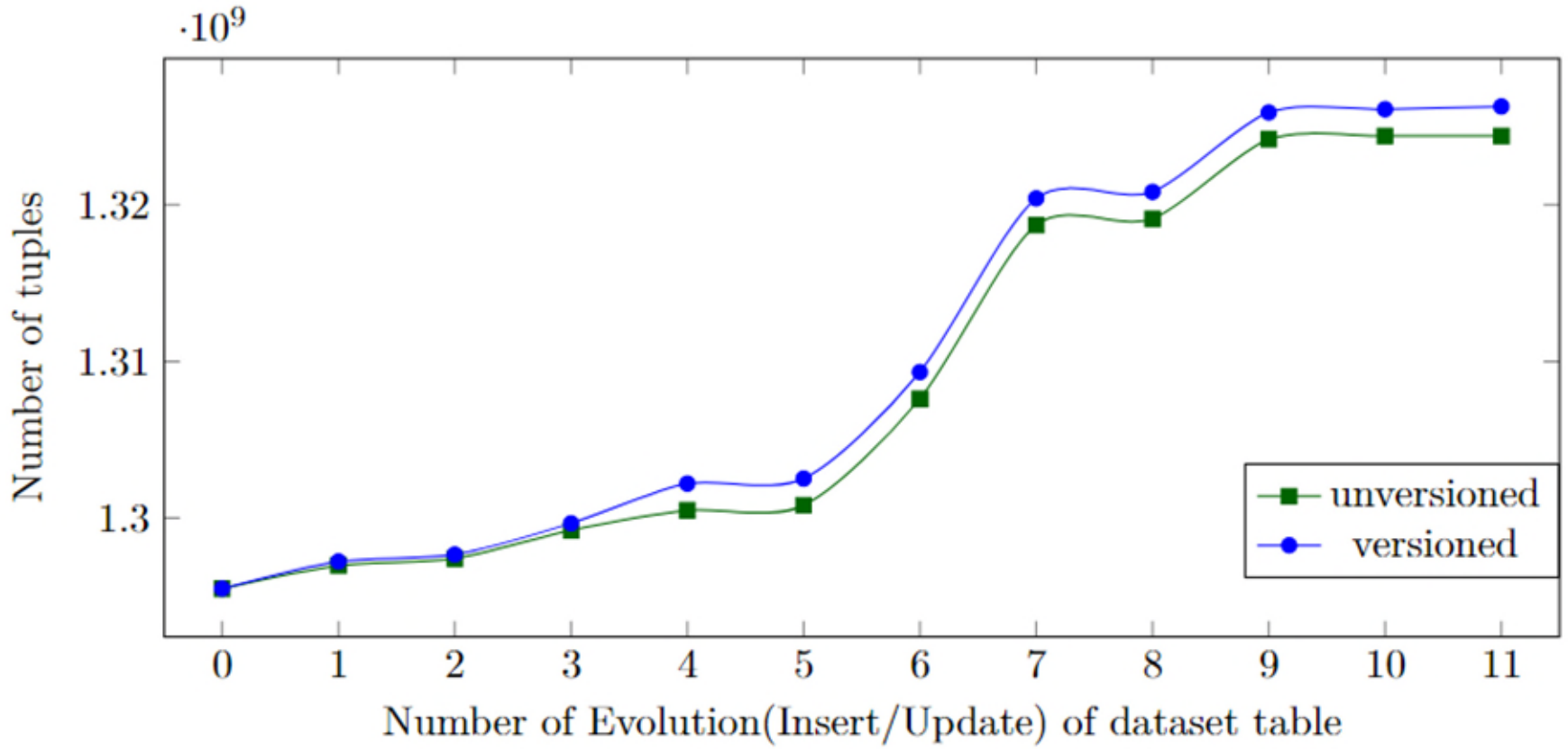


Evaluation Results

- Iterative Integration of new data of 10 selected networks of various size
- Execution of queries after each update
- Measurement of storage increase in absolute and relative size
- Measurement of query-execution times and update/insert times
- Verification of query re-execution correctness

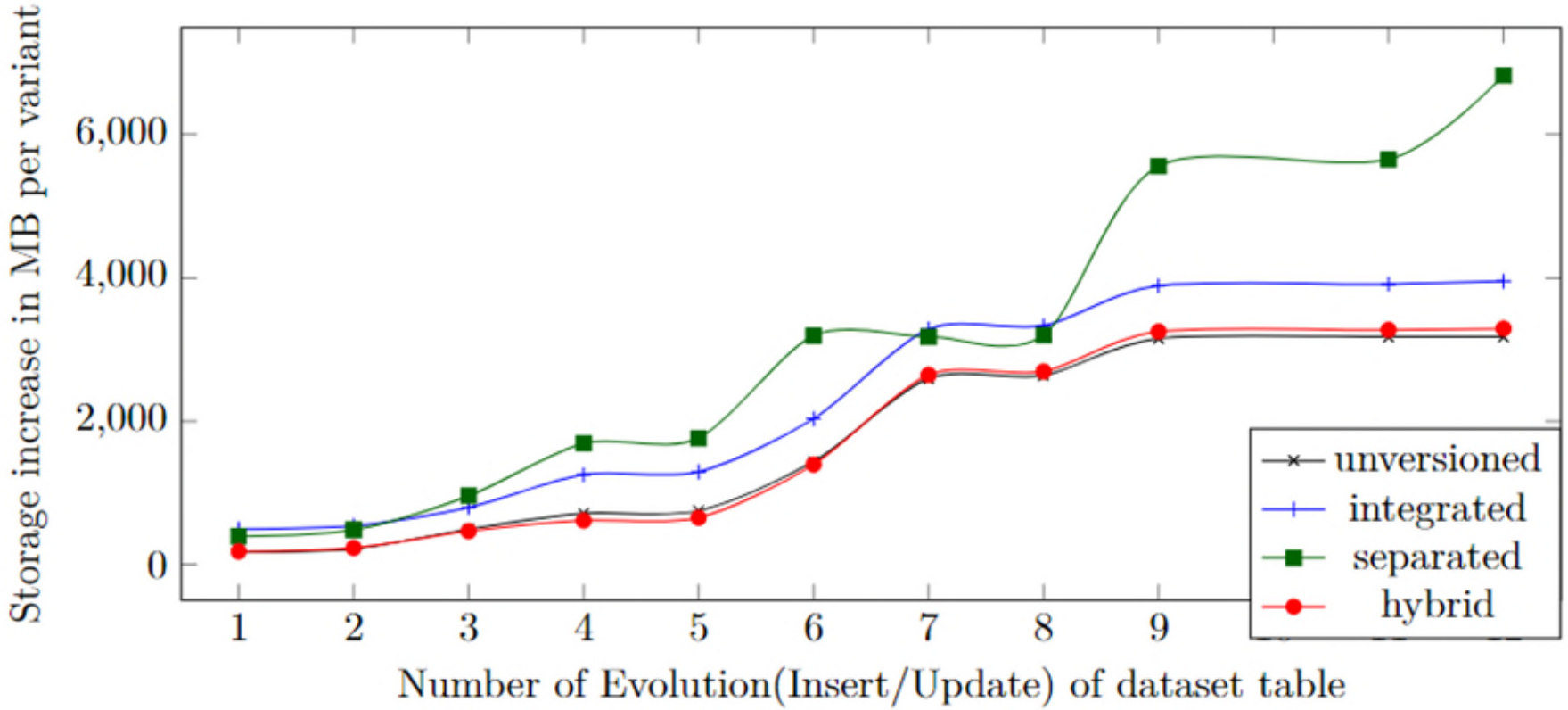
[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Cummulativ increase of rows

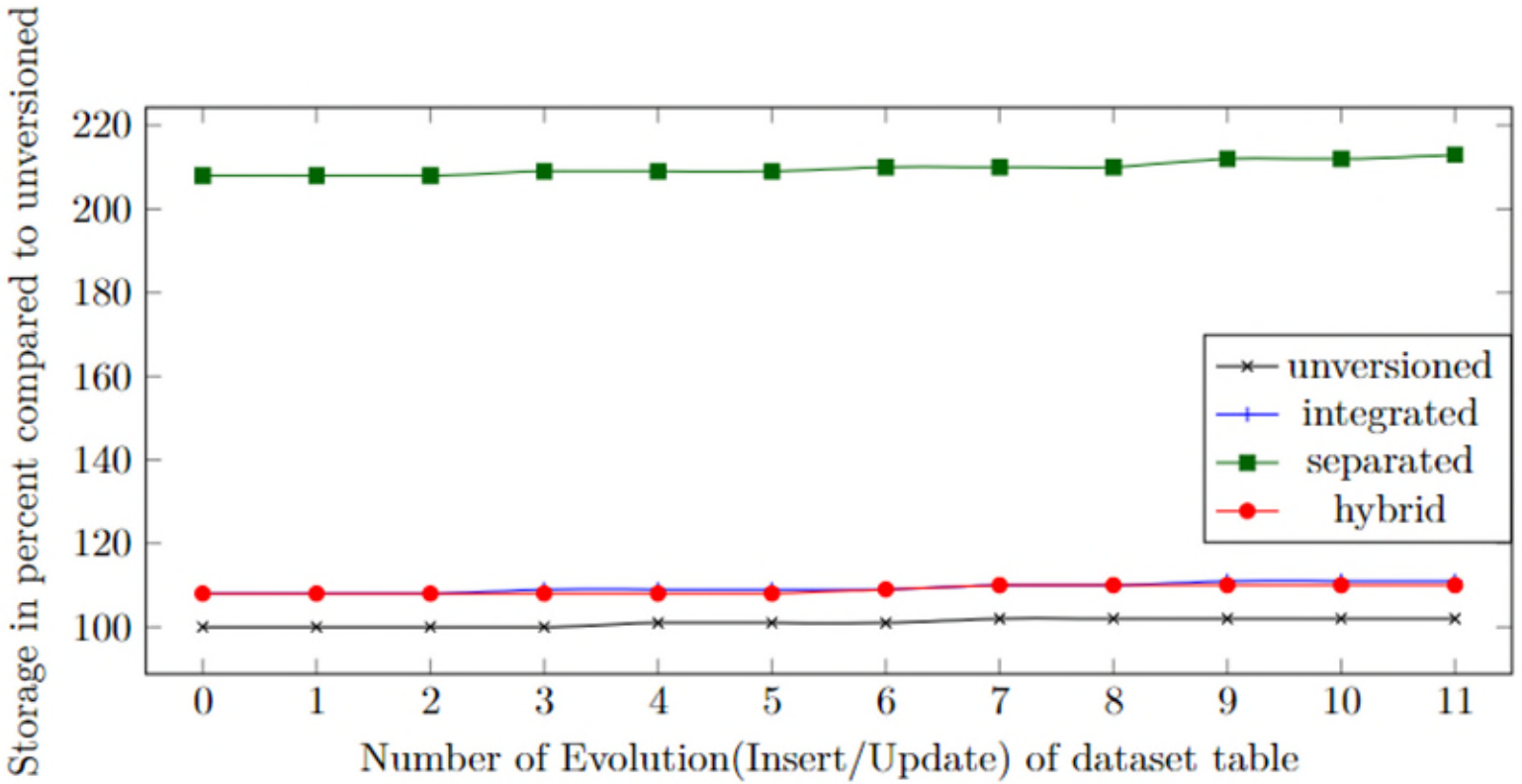


[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Increase of Storage per Evolution



Storage in Percent compared to no-versioning



Evaluation results

Versioning method	Versioned db	Implementation	Storage space increase	Performance Query	Performance Re-execution	Future amendments to the db
Integrated: **→ public db (all)**	Original db	Easiest (creation of "valid from" and "valid to" columns in each table)	Increase (two new columns added to each table) - 32 bytes per tuple (> +8 %)	On versioned public db - 3 extra constraints – the bigger the database the more run time necessary → Average run time 17:37 min	On versioned public table → Average run time 1:23 min	Complex to handle since the full db is versioned
Separated: → **public db (most recent)** → **history db (all)**	History db	Some complexity (creation of versioned history db – picture of all)	Stores every tuple twice (original and history) + stores changes in history table - 32 bytes per tuple (+ >73%)	On unversioned public database - query remains unchanged → Average run time 17:11 min	On versioned history table with → Average run time 1:23 min	Easier to handle due to the separation (unversioned public and versioned history table)
Hybrid: → **public db (most recent)** → **history db (old)**	Both dbs	Most complex (versioning on both tables and split up into most recent and old)	Split of original table in two + two new columns added - 16 bytes per tuple (+> 4%)	On versioned public db - Query remains unchanged → Average run time 15:06 min	On versioned public and versioned history table →Average run time 19:02 min	Very complex to handle since the db is split up in the versioned public db and the versioned history db

Most complex/ not good in comparison

Best in comparison

[1,5] TU Wien: [1]climers.geo.tuwien.ac.at & [5]Informatics Department
[2] AWST: www.awst.at
[3] CESBIO: www.cesbio.cnrs.fr
[4] ESA: https://earth.esa.int/eogateway

Going eScience

- Interdisciplinary Research Community conference

- 9 – 13 October, 2023 (Cyprus)
- Submission of paper
- https://www.escience-conference.org/2023/

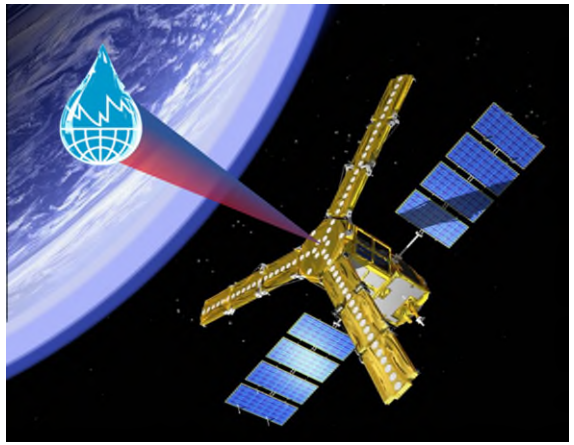
References

[1] Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Pröll. “Identification of reproducible subsets for data citation, sharing and re-use”. Bulletin of the IEEE Technical Committe on Digital Libraries (TCDL), 12(1), May 2016.

[2] Andreas Rauber, Bernhard Gößwein, Carlo Maria Zwölf, Chris Schubert, Florian Wörister, James Duncan, Katharina Flicker, Koji Zettsu, Kristof Meixner, Leslie D. McIntosh, Reyna Jenkyns, Stefan Pröll, Tomasz Miksa, and Mark A. Parsons. “Precisely and persistently identifying and citing arbitrary subsets of dynamic data”. Harvard Data Science Review, 3(4), 11 2021. https://hdsr.mitpress.mit.edu/pub/si7wzxxa.

[3] M. Wilkinson, M. Dumontier, I. Aalbersberg et al. “The FAIR Guiding Principles for scientific data management and stewardship”. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

ISMN in house issuing of DOIs



Thank you for your attention!

https://project-frm4sm.geo.tuwien.ac.at/

Contact

Contact FRM4SM project: irene.himmelbauer@geo.tuwien.ac.at; Alexander.gruber@geo.tuwien.ac.at

Contact Data Citation experts: Tomasz.miksa@tuwien.ac.at; moritz.Staudinger@tuwien.ac.at

Irene Himmelbauer[1], Moritz Staudinger[5], Tobias Hajszan[5], Tomasz Miksa[5], Andreas Rauber[5], Daniel Aberer[1], Alexander Gruber[1], Wolfgang Preimesberger[1], Pietro Stradiotti[1], Wouter Dorigo[1], Monika Tercjak[2], Alexander Boresch[2], Arnaud Mialon[3], Francois Gibon[3], Philippe Richeaume[3] , Yann Kerr[3], Raul Diez Garcia[4], Raffaele Crapolicchio[4], Roberto Sabia[4], Klaus Skipal[4], Philippe Goryl[4]