

BONSAmurai meeting

4:00pm CET - 27/09/18 via Zoom

Minutes of the meeting

Present: Agneta Ghose, Chris Mutel, Konstantin Stadler, Massimo Pizzol, Stefano Merciai, Roman Sacchi, Tomas Navarrete, Michele De Rosa (minutes), Bo Weidema (minutes editing)

1) Bo shared some news on classifications:

- [SDMX registry](#), a rapidly developing framework for statistical data publishing, closely related to the RDF data cube;
- GS1 [Global Product Classification \(GPC\)](#) is now available as an open hierarchical and highly detailed classification, which seems the natural starting point for the "product" part of our "flow-object" dimension, linking also to CPC. Could also be the starting point for disaggregating the ISIC classification.

2) Bo informed that EXIOBASE Version 3.3 is now available with a CC-SA license, so that we can use this as a temporary foundation for our further experimentation towards a pilot version of the BONSAI database. The Supply-Use tables are available for the hybrid version, which is the most relevant for BONSAI. Konstantin would investigate if Supply-Use tables will also be made available for the monetary tables. The 3.4 version will come soon. However, for testing, it is not so important to use the latest version.

3) Bo showed the wiki page on data format: <https://github.com/BONSAMURAI/bonsai/wiki/Data-Storage#specify-minimum-core-data-and-metadata-formats> where more details have been added during the last months.

Chris Mutel announced that he had some issues and questions but thought it might be better to formulate these in writing. It was suggested to use the 'issues' tab on GitHub for discussing wiki's items and for advancing proposals for contributions, because the wiki itself is not ideal for that purpose, although it stores the chronology of modifications.

The following comments were made on the BONSAI wiki text:

- Unit: Bo recommended the OM ontology based on a recent review and comparison. For currencies, units must be supplemented by a time stamp and information on purchasing power correction and other adjustments (it seems that this is currently not in any existing unit ontologies, so we have to make our own addition here). Bo asked if this should it be in the unit name or separate? The feeling was that it would be best to keep as separate property information for the units. The question was raised how SDMX deals with basic and purchaser price? It seems that it is currently just listed as additional information but it is not clear how it is used in practice.

- From our basic philosophy of limiting the required data fields to a minimum, Bo said that uncertainty should not be a required field. Bo suggested that uncertainty could be adopted from the OLCA format. Chris Mutel will provide an alternative proposal.
- Chris Mutel and Bo will also work together on the geolocation format.

There was a general agreement on the presented suggestions on the data format; it needs to be tested out now, and spelled out in e.g. JSON-LD.

Tomas was suggested that Cyril Francois may have done a conversion of SPOLD to RDF already, so we should avoid re-doing this to the extent we could reuse that. Bo mentioned that he had asked Cyril to comment on the wiki text on the database format, but that Cyril has not yet had time to do so.

4) As RDF Database management system, Bo suggested to use Apache Jena (most widely used) rather than CKAN, which had been previously proposed. These are in two different languages though. CKAN is in python while Apache Jena is in Java, so the choice could also depend on what language our contributors feel most at home with. Initially there is no problem in experimenting with both – the important thing is to have the database in RDF – how it is stored can also be decided later.

We have been granted space on the Aalborg university server for hosting a database for testing purposes.

The next task is to make an RDF-version of EXIOBASE v3, using the data format discussed above, and placing this on the joint server for further experimenting. It was suggested that a conversion back to the original sql (or excel) format would then be a good test.

5) Having the database in RDF would then allow us to test the integration of process datasets into the "raw data" part of the database. Actual numerical operations on the database, such as balancing and matrix inversion, may best be done when the modified RDF data are translated back into a traditional SQL (relational) database format. Bo confirmed that the overall final goal is still to have a product footprint calculator, but that the integration test (IO-data and process data) in RDF is a first milestone.

Chris suggested that this and the previous task may be the object of an in-person meeting/hackathon since none may feel comfortable to do it alone. A proposal on an in-person hackathon workshop with key members of the group.

6) Funding and planning for a physical "Bonsamurai" workshop

An in-person hackathon workshop with key developer's members of the group was proposed. Key members may be able to fund their own travel and accommodation. It may be possible to have some funds from Aalborg University for the meeting (not necessarily in Aalborg). For some participants it may be easier to cover costs if the meeting is organized as a scientific stay at a university institution.

As critical members that should be asked to participate in a doodle to identify a potential meeting date we identified: Chris Mutel, Cyril Francois, Tomas Navarette, Bo, Konstatin Stadler. Chris and Bo would make a more detailed proposal for an agenda for the hackathon. If anyone have contacts that have more experience with coding for RDF and should be invited, please suggest to Bo.

It was suggested that it could be advantageous to present the RDF data in a form that is easier to digest for people more used to SQL databases.

The meeting was closed after 1 hour and 10 minutes.