Release Note for the THOR Dashboard

Document Information

Date: 12/07/2017

Authors: Sünje Dallmeier-Tiessen (CERN, orcid.org/0000-0002-6137-2348)
Robin Dasler (CERN, orcid.org/0000-0002-4695-7874)
Anna Trzcinska (CERN, orcid.org/0000-0002-5601-2479)
Ioannis Tsanaktsidis (CERN, orcid.org/0000-0002-1567-3676)

Reviewers: Tom Demeranville (ORCID)
Kristian Garza (DataCite)

Abstract: The THOR dashboard was first released in March 2016. Discussions with stakeholders and reviewers, as well as changes in available resources and project priorities have prompted a series of updates. Suggestions were incorporated, engineering was improved to make the dashboard more scalable, robust, and easier to maintain. This report summarises the changes included through June 2017.

DOI: 10.5281/zenodo.826163

This work was supported by the THOR Project. The THOR project is funded by the European Union under H2020-EINFRA-2014-2 (Grant Agreement number 654039). The following report is based on a deliverable submitted to the European Union on 30 June 2017.

Visit http://project-thor.eu for more information.

Executive Summary

Since the release of the THOR dashboard in March 2016, discussions with stakeholders and reviewers have highlighted additional data and options that could be included to make it more comprehensive for the wider community. This, along with changes in available resources and shifting interests of the THOR team, have led to a series of updates during the second year of THOR. Most notably, this resulted in the incorporation of some basic data from Crossref, as well as operational improvements to make the dashboard more scalable and robust.

This short report outlines what has been done, what considerations were taken into account, and how others can benefit from this work. It also highlights challenges encountered and lessons learnt along the way, including design limitations, the creation of tutorials for creating similar services, and the issue of scope creep.

The dashboard is available to view on the THOR project website[1], and the code (in two parts) is available on github[2,3].

[1] http://project-thor.eu

[2] https://github.com/thor-project/dashboard

[3] https://github.com/thor-project/data-harvester

Contents

1 Background

The sustainability branch of THOR was originally tasked with observing the evolution of the persistent identifier (PID) landscape to serve as a contextual basis for evaluating the project. In order to facilitate this observation, the team's task during the first six months of the project was to devise a set of metrics. This set of metrics evolved into a dashboard, a web application for aggregating metrics from the two THOR infrastructural partners. This dashboard was released in March 2016 and can be found on the project website[4].

The dashboard originally had a single simple use case: observe the evolution of the PID landscape. The complete requirements, engineering process and design philosophy were described in 'THOR: Metrics and Tools' (Dasler, 2016). After the dashboard's release, discussions with stakeholders and reviewers have highlighted additional data and options that could be included to make it more comprehensive for the wider community. Taking into account these suggestions, as well as the changes in available resources and the shifting interests of the THOR team over the course of the project, the dashboard has since been updated. This short report, coinciding with the second anniversary of the THOR project, outlines what has been done, what considerations were taken into account, and how others can profit from this work.

2 Technical Details

A number of technical fixes and improvements were implemented to improve the operational efficiency of the THOR dashboard and to align it with the common development practices now in use for other CERN Scientific Information Service (SIS) products. These were made up of both 'under the hood' performance improvements and improvements in functionality for the dashboard.

2.1 Improvements in Operational Efficiency

As part of a more general operational streamlining effort within the CERN SIS team, the THOR dashboard was moved to new servers. This provided an opportunity to make changes to its underlying structure, configuration and development workflow.

The development of the dashboard is now split across development ('dev'), testing ('qa'), and master ('master') git branches to allow easier testing of new features prior to launch. New features are developed on the dev branch, pushed to qa for testing, and sent to master once finalised

Underneath, separate production and development servers were configured with Puppet configuration management[5]. The dev branch pushes to the development server, and the master branch pushes to the production server, while the qa branch is for intermediate testing and quality assurance.

On its original server, the dashboard used a local database for storing events information. When it was moved to new servers, that database was migrated to a centrally managed MySQL database through the

[4] http://project-thor.eu

[5] https://puppet.com/

CERN DB On Demand service[6]. This service provides support for updates, backup and recovery, and service continuity, freeing up developer time that would otherwise have been spent on these maintenance activities. Otherwise, the functionality of the database is the same. Others wishing to deploy their own copy of the dashboard do not need access to the CERN service.

2.2 Improvements to Dashboard Functionality

2.2.1 Adding Crossref Metrics

The most visible change to the dashboard since its launch is the addition of metrics for Crossref DOIs (see Figure 1). This addition was recommended by reviewers during the THOR first-year review. While Crossref is not a THOR partner, the organisation is a major DOI provider in scholarly communications. Including metrics for Crossref DOIs would complement the existing metrics for DataCite DOIs and ORCID iDs, and complete the article–data–researcher triangle.

Given the scope of the rest of the dashboard, it was determined that the most relevant information available via the Crossref API was the number of DOIs minted monthly and the cumulative total of Crossref records that contain an ORCID iD. This information fits logically with the other information available in the dashboard and requires no local data processing. Crossref provides additional information about its DOIs, but other information was deemed either irrelevant for THOR purposes (such as number of records with licences) or impractical to display (such as the over 8,000 member organisations).

The data in the dashboard is populated through harvester modules[7], with a corresponding harvester module for each service from which data is pulled (that is, DataCite and ORCID). An additional harvester was created to pull selected data from the Crossref API[8].

A separate dashboard display package[9], built with the dc.js javascript charting library[10], handles the display of the information pulled from each harvester. An additional display module for displaying the harvested Crossref data was added to this package[11].

2.2.2 Automating Harvesting

In the first release, harvesting of the information occurred monthly. This was because some of the partner statistics were only available on a monthly basis, and more frequent updates would have yielded a negligible increase in utility. The monthly harvest was run manually. With the server move, we took the opportunity to schedule automated processes for the monthly updates to minimise operational effort. The code for the harvester is in the separate harvester package on GitHub[12], and the updates are now run by a server-side chronjob.

---

[6] https://cern.ch/DBOnDemand

[7] https://github.com/thor-project/data-harvester

[8] https://github.com/thor-project/data-harvester/commit/7c861791941197d635465f254ea09eb0565b740e

[9] https://github.com/thor-project/dashboard

[10] https://dc-js.github.io/dc.js/

[11] https://github.com/thor-project/dashboard/commit/e7f46f5152e63569ea70cbc19c2fa55cec5a8b48

[12] https://github.com/thor-project/data-harvester

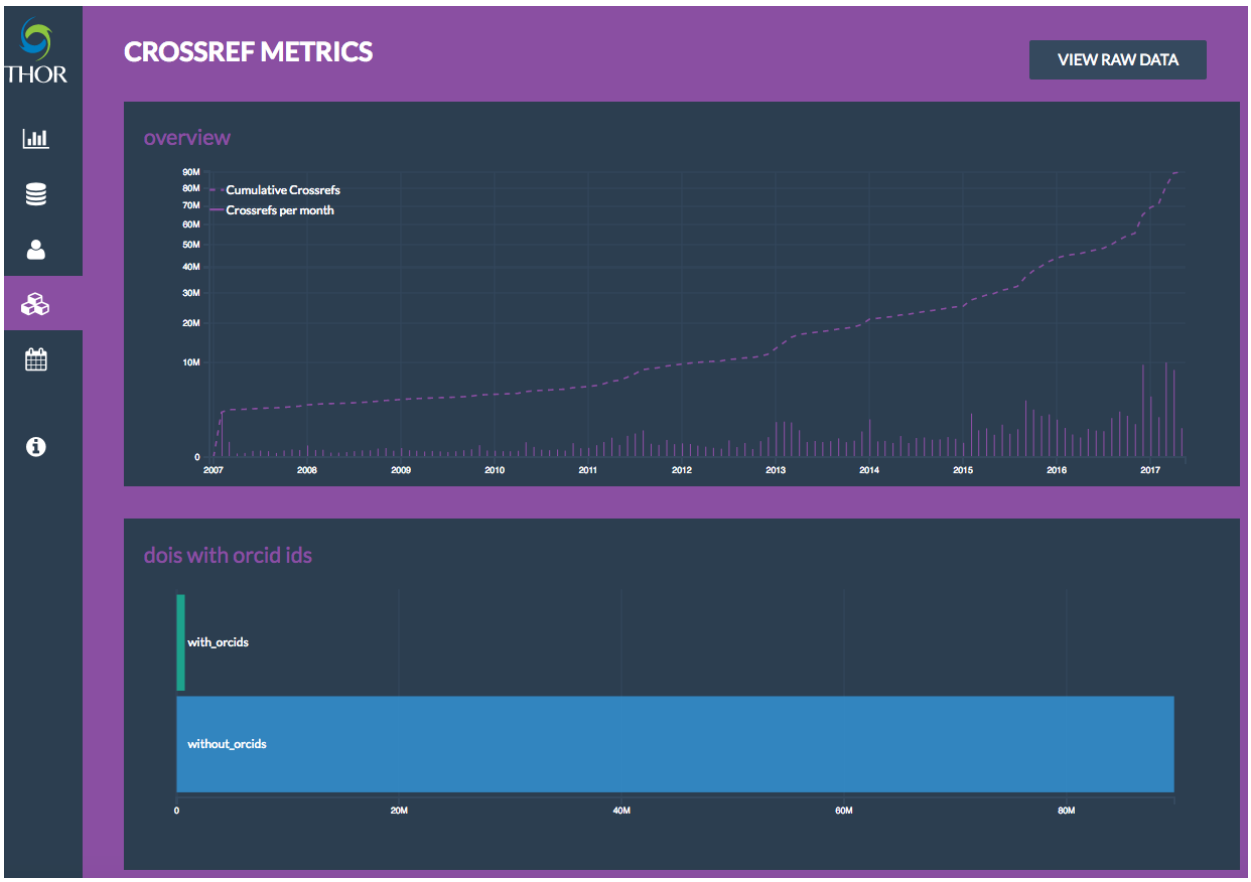

Figure 1: Display of Crossref metrics in the dashboard

2.2.3 Usage statistics

The original dashboard did not include a way to track usage statistics. The initial design assumption was that the dashboard, though perhaps academically interesting to others, would be primarily useful to THOR partners, so the usage of the dashboard did not seem relevant to our outreach metrics. As it has become apparent that this is not the case, Google Analytics was added to the dashboard, so that usage statistics from the later stages of the THOR project can be included at the project's final review.

2.2.4 General Content Updates

In addition to the more complex performance and feature updates already mentioned, a few minor content updates were also completed:

- Calendar display was updated to display the most recent three years of events pulled from the events list (see Figure 2).
- Labels were changed to match preferred partner wording (for example, ORCID iDs vs. ORCIDs).

Figure 2: The dashboard's event calendar, which displays THOR events

3 Reuse of the Dashboard

In September 2015, the sustainability team held a set of focus groups to gather feedback on the dashboard prototype (Dasler, 2016). At that time, people from various stakeholder groups expressed enthusiasm for the dashboard and an interest in developing similar resources customised for their specific institutional use cases.

The code underlying the dashboard has been available publicly on github[13] since the beginning of its development (see Figure 3), enabling others to reuse the code or to be inspired to develop similar tools. There is also a supporting documentation page included as part of the THOR Knowledge Hub[14]. However, both are insufficient to enable others to get started. In order to provide support necessary to jumpstart community development around projects like the dashboard, it is beneficial to provide hands-on guidance and experience.

In the first bootcamp organised as part of the THOR project, which was held in Madrid, we included the dashboard as part of an introductory technical development track (Chen et al., 2017). The motivation was twofold: to meet the expressed interest and to encourage the dissemination of THOR outputs. Tutorials[15,16] were developed to guide participants with any level of technical background through the

[13] https://github.com/thor-project/dashboard and https://github.com/thor-project/data-harvester

[14] https://project-thor.readme.io/docs/dashboard

[15] https://thor-project.github.io/dashboard-tutorial/

[16] https://github.com/thor-project/dashboard-tutorial

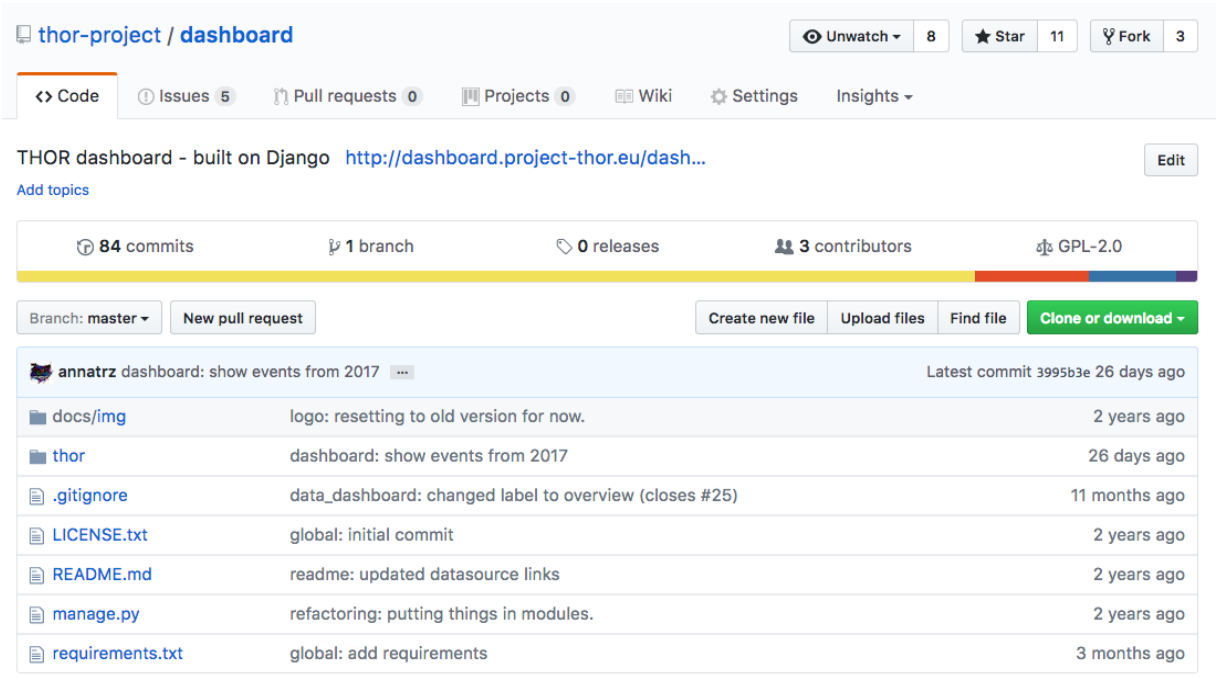

Figure 3: THOR dashboard github repository: https://github.com/thor-project/dashboard

development process of a similar tool. Even participants with no coding skills were able to quickly get an understanding of how they could profit from the publicly available datasets and build a tool to assess or observe global developments in Open Science. Having finished the tutorial, participants have the capability to tailor such a dashboard to include, for example, data from their own institution (such as their institutional repositories). This experience underlines the need to further the creation and dissemination of robust and easy to implement tools, such as this dashboard.

4 Challenges and Lessons Learnt

The process of maintaining the dashboard throughout the project presented some challenges around both its content and implementation. The initial design challenges and lessons learnt are presented in 'THOR: Metrics and Tools' (Dasler 2016), but additional lessons emerged throughout the course of the project.

One particular challenge that became apparent over the course of the project concerned the integration of the THOR event calendar. Originally, the hope was that plotting the calendar events alongside the PIDs in the overview chart would enable one to see possible correlations between THOR events and increases in the adoption of the identifiers concerned. Perhaps unsurprisingly, the continued growth of each PID type has been relatively steady over the course of the project, with few perceptible changes in PID assignment around the time of THOR events. Further, any fluctuations that do appear in the data are difficult to attribute to THOR. Finally, the range of valuable impacts that THOR events could have are not well represented by the sole criterion of PID assignment.

These issues, in part, led to the proposal of a service adoption gap analysis study[17], which the THOR sustainability and outreach teams took on as an additional activity over the project's second year. As the gap analysis shows, a more in-depth and detailed analysis is needed in order to correlate activities with the range of impacts that are useful for informing further outreach efforts. Within the gap analysis, this in-depth analysis was only possible due to the intense data processing focused on disciplines and regions, which is not feasible to do on the fly for these 'living' datasets.

Another challenge that we have encountered over the course of maintaining the dashboard has been that of scope creep: balancing the temptation to add more data and visualisations to the dashboard with the realisation of the goals of the project. While it can be tempting to continuously add components to the dashboard, we must keep in mind which components might serve the evaluative purpose of the dashboard and which components present the underlying data in a way that does not obscure their meaning. This is especially important given the potential added maintenance costs or sustainability concerns of additional components. Scope creep is not an unexpected challenge in this context, as it is always challenging to strike a balance between innovation and sustainable services; but the endurance of this challenge was not anticipated when we were originally designing the dashboard.

Fortunately, a solution to this challenge is presented by the open nature of THOR. As the development of the dashboard and the tutorials developed for the bootcamp are open source, it can be left up to the community to build and extend dashboards that meet their own specific and diverse data visualisation needs. THOR has built the foundations for these activities, and its open source principles can help leverage these innovations for the broader community. Our experience with the bootcamp tutorial showed that there is considerable interest to further these kinds of activities.

Related to the scope creep challenge is the piecemeal challenge. Feedback on the dashboard revealed considerable interest in data visualisations to enable easy assessment of current developments in specific e-infrastructures, specifically with the ability to surface details of individual PIDs. The dashboard is by design an aggregator, harvesting information from multiple sources for the benefit of displaying it in a visually useful way, in one place. By necessity, the dashboard's design focuses on a high-level view of the registries concerned. For future work, it would be interesting to add another layer to study individual PIDs in a more focused way. This might be relevant for innovations like Scholix[18] or DataCite[19]/Crossref[20] Event Data. Such an expansion requires further discussion, as it would require further data processing post-harvest and would impact display design.

Finally, a minor lesson learned is not to underestimate the value that this type of visualisation project has for the wider PID community. The dashboard did not originally include any analytics on its own use, as it was intended primarily to track trends of interest to the project, and it was therefore assumed that external usage would be low. However, the interest external stakeholders have expressed in the dashboard has been greater than anticipated.

[17] Publication forthcoming on THOR website, http://project-thor.eu

[18] http://www.scholix.org/

[19] https://eventdata.datacite.org/

[20] https://www.crossref.org/services/event-data/

5 Future of the Dashboard

The dashboard is currently hosted at CERN on trusted and sustainable CERN resources. However, THOR is a project with a limited lifespan, so perpetual maintenance of the dashboard is untenable. Given the interest in developing further features in this type of tool, it is being discussed whether and how the dashboard could become a utility that is continued in organisations like DataCite or initiatives like Scholix[21].

6 Conclusions

Over the life of the THOR project, the dashboard has proven to be a useful exercise in assessing the state of PID metrics, and it has sparked a great deal of interest among THOR's stakeholders. The improvements in performance and functionality described in this document have served to further enhance its value, despite several challenges. The open nature of the development has laid the necessary groundwork for the PID community to take up and drive projects of this kind. Though THOR has a limited lifetime, there is potential for the dashboard, or other tools based on it, to be created and expanded beyond the project's lifetime.

7 References

Dasler, Robin. (2016). THOR: Metrics and Tools. *Zenodo*. doi: doi.org/10.5281/zenodo.46761

Chen, Xiaoli, Dasler, Robin, Rueda, Laura and Maaike Duine. (2017). Capacity Building Report – Bootcamps. Internal Report.

[21] http://www.scholix.org/

Appendix A: Terminology

Additional terms are defined below:

Term	Definition
API	Application programming interface
CERN	CERN, the European Organization for Nuclear Research, is one of the world's largest centres for scientific research. http://home.cern/
CERN SIS	CERN Scientific Information Service
DataCite	An organisation that develops and supports methods to locate, identify and cite data and other research objects. Specifically, DataCite develops and supports the standards behind persistent identifiers for data, and the members assign them. See https://www.datacite.org
DOI	Digital Object Identifier
ID	Identifier
ORCID	An organisation that creates and maintains a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. See http://orcid.org
PID	Persistent Identifier
Scholix	High level interoperability framework for exchanging information about the links between scholarly literature and data. See http://www.scholix.org/
THOR	Technical and Human Infrastructure for Open Research. See https://project-thor.eu/

Appendix B: Project Summary

The **THOR** project establishes a sustainable international e-infrastructure for persistent identifiers that enables long-term access to critical information about the life cycle of research projects. It enables seamless integration between articles, data, and researcher information creating a wealth of open resources. This will result in reduced duplication, economies of scale, richer research services, and opportunities for innovation.

The project has four concrete aims:

1. Establishing interoperability
2. Integrating services
3. Building capacity
4. Achieving sustainability

The project will meet these aims by defining relations between contributors, research artefacts (including data), and organisations. We will incorporate these relationships into the ORCID and DataCite systems. We will also expand existing linkages between different types of identifiers and versions of artefacts to improve interoperability across platforms and integrate ORCID iDs into production systems for article and data submission services in pilot communities and beyond.

The consortium will develop systems to embed new PID resolution techniques into existing services to support seamless direct access to artefacts, and in particular data. We will create services to allow associations between datasets, articles, contributors and organisations at the time of submission. Building on these, we will deliver the means to integrate trans-disciplinary PID services in community-specific platforms, focussing on cross-linking, claiming mechanisms and data citation (guided by the FORCE 11 data citation principles[22]).

For more information, visit http://project-thor.eu or contact info@project-thor.eu.

[22] https://www.force11.org/group/joint-declaration-data-citation-principles-final