UNISDR Global Assessment Report - Current and Emerging Data and Compute Challenges

Nils Gentschen Felde, Mabel Fraume, Matti Heikkurinen, Dieter Kranzlmüller, Julio Serje

UNISDR Global Assessment Report

Current and Emerging Data and Compute Challenges

Nils gentschen Felde[1], Mabel Cristina Marulanda Fraume[2] Matti Heikkurinen[1], Dieter Kranzlmüller[3], Julio Serje[2]

[1]Ludwig-Maximilians-Universität München (LMU), Munich Germany
{felde,heikku}@nm.ifi.lmu.de
[2]The United Nations Office for Disaster Risk Reduction (UNISDR), Geneva, Switzerland
{marulandafraume,serje}@un.org
[3]Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Garching bei Münich, Germany
Dieter.Kranzlmueller@lrz.de

Abstract. This paper discusses the data and compute challenges of the global collaboration producing the UNISDR Global Assessment Report on Disaster Risk Reduction. The assessment produces estimates – such as the "Probable Maximum Loss" – of the annual disaster losses due to natural hazards. The data is produced by multi-disciplinary teams in different organisations and countries that need to manage their compute and data challenges in a coherent and consistent manner.

The compute challenge can be broken down into two phases: hazard modelling and loss calculation. The modelling is based on production of datasets describing flood, earthquake, storm etc. scenarios, typically thousands or tens of thousands scenarios per country. Transferring these datasets for the loss calculation presents a challenge – already at the current resolution used in the simulations. The loss calculation analyses the likely impact of these scenarios based on the location of the population and assets, and the risk reduction mechanisms (such as early warning systems or zoning regulations) in place. As the loss calculation is the final stage in the production of the assessment report, the algorithms were optimised to minimise risks of delays. This also paves the way for a more dynamic assessment approach, allowing refining national or regional analysis "on demand".

The most obvious driver of the future compute and data challenges will be the increased spatial resolution of the assessment that is needed to more accurately reflect the impact of natural disasters. However, the changes in the production model mentioned above and changing policy frameworks will also play a role. In parallel to these developments, aligning the current community engagement approaches (such as the open data portal) with the internal data management practices holds considerable promise for further improvements.

Keywords. Hazard and loss modelling, probabilistic modelling, disaster risk, distributed data management

adfa, p. 1, 2011.

1 Introduction

This paper describes the data and compute challenges related to the production of the biennial Global Assessment Reports (GAR) [1], key documents providing high-level overviews of the status of the disaster risk reduction activities on the global level. The production of the data these documents are based on – hazard scenarios, exposure information and vulnerability modelling – is performed by distributed collaborations and coordinated by UN Office for Disaster Risk Reduction (UNISDR). The risk calculation that provides estimates of the likely annual losses due to natural disasters represents a time-critical challenge, both in terms of organising the necessary computational processes to produce and verify the results, and in terms of managing the data sets in a consistent way. In parallel to the GAR-specific analysis, making the modelling data available as an open data service could support numerous additional research activities. These issues and goals form the context of the ongoing collaboration between UNISDR, LMU and LRZ.

While the type of large-scale disaster risk modelling GAR represents is most likely unique, we can see similarities with overall process and organisation as well as technical approaches with certain initiatives in other research domains. For example, the overall organisational structure resembles the approach used by global High-Energy Physics (HEP) collaborations. However, managing the complex, interdependent evolution of the interfaces (both physical and software ones) between thousands of components in a typical HEP project has necessitating developing relatively rigid and formalised organisational processes. This makes most of the tools developed to support HEP collaborations (ranging from document management systems - such as EDMS [2] - to global data/compute systems such as Worldwide LHC Computing Grid [3]) developed for HEP collaborations not optimal for the GAR process.

Perhaps the closest analogue can be found from the earth observation domain. The Group on Earth Observations (GEO) has launched the GEO-DARMA [4] initiative with a goal of bringing earth observation data into disaster risk management. The effort builds on earlier initiatives focusing on specific hazards (floods, volcanoes) and aims to extend the focus from supporting the immediate, acute response to supporting preparedness and risk reduction. However, at the time of writing the initiative is still in its early stages.

2 Global Assessment Report - GAR

2.1 Background

UNISDR was established in 1999 and its role in the UN system is to serve as the focal point for disaster risk reduction activities to ensure coordination and synergies between UN organisations, and regional and national activities. The two major UN policy documents bringing all this guidance together into top-level policy documents are the 2005 Hyogo Framework for Action [5] and the 2015 Sendai Framework for Disaster Risk Reduction [6]. The GAR process played a key role in implementing the Hyogo

Framework and in the preparations of the Sendai Framework. It provided concrete data and examples of how the policies implemented (or to be implemented), changing natural conditions (e.g. climate change), population movements and major infrastructure projects are influencing the likely consequences (lives lost and direct economic losses) of natural hazards.

On the abstract level, the Global Risk Assessment processes of GAR 13 and GAR 15 are based on building three datasets used to generate risk metrics such as the "Loss Exceedance Curve" (LEC), the "Average Annual Loss" (AAL) and the Probable Maximum Loss (PML) plot the for the different hazards considered for each country. The data necessary for performing the global risk assessment of the GAR15report are:

- Hazard data, consisting of groups of simulated scenarios for each of the natural hazards for each of the natural hazard (earthquake, tsunami, riverine flood, cyclonic wind, storm surge and so on) used in the analysis. Each set of simulated scenarios must comply with the certain key requirements, such as being mutually exclusive, collectively exhaustive and having an annual frequency of occurrence associated with them.
- Exposure data, describing each exposed asset with a set of attributes such as their geographical location, structural characteristics, construction material type, economic value (among others).
- Vulnerability data, characterising the exposed asset with a set of attributes describing their relevant characteristics that determine how sensitive they are to different hazards at different intensities.

Current GAR information linked to other datasets allow further evaluations and analysis. For example, risk associated to hydrometeorological hazards are strongly influenced by climate change, hence the IPCC data IPCC [7] data and reports used as an input for simulation of new hazard scenarios considering climate change. Similarly, the exposure data used includes contributions collected using crowdsourcing approach based on OpenStreetmap [8] and the vulnerability data can be improved by counting with better information of exposed assets. Additional data sources, such as OECD macroeconomic data, are used to conceptualise the disaster risk metrics. These dependencies are a partial rationale for the relatively frequent releases of the GAR.

The key output of the analysis are the country-level summaries presented in the Global Assessment report (a typical view of the online version is presented in Figure 1). This distils a complex and multi-faceted analysis into a summary with few key indicators that are suitable for steering policy-level decision making in the UN member states.

As a result, the environmental modelling behind GAR can have a major societal impact. While not legally binding, the GAR recommendations have an impact on national legislation and e.g. zoning decisions – both areas that unavoidably have an impact on economic situations and prospects of both public and private sector entities. In the long-term it has been shown that most of the investments in risk reduction are "profitable" in the sense that investments will eventually prevent direct economic losses that would have been several times higher than the money spent on protection (as an example, analysis of government-funded flood defence schemes in the UK showed an average benefit–cost ratio or 9.5:1 [9]). However, statistically major natural hazards usually

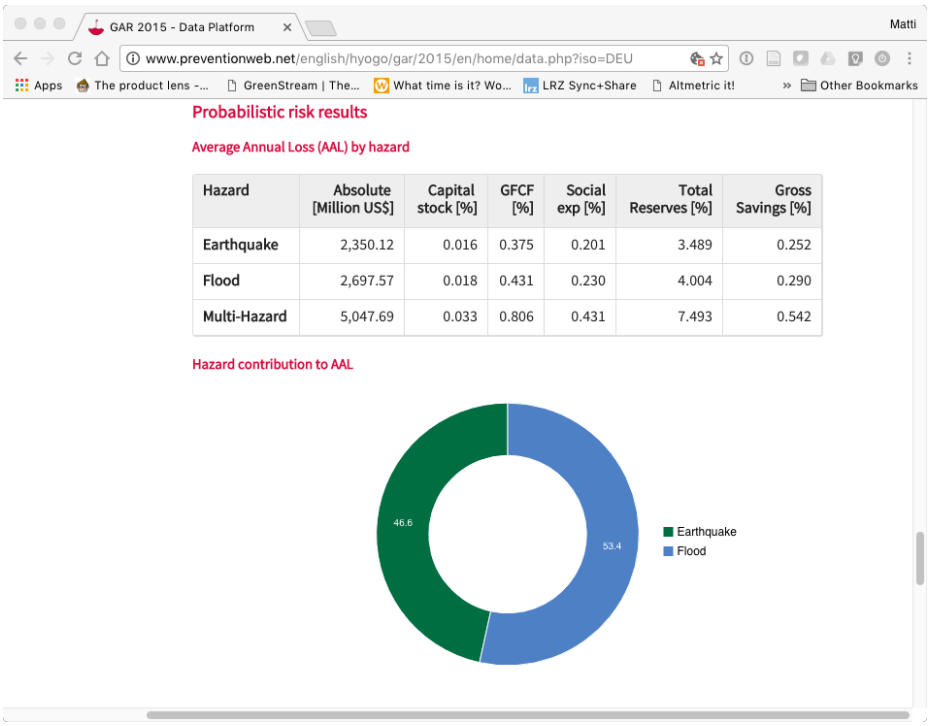
have several decades between each occurrence, complicating the benefit-cost calculations. Justifying the immediate costs (loss of revenue or increased tax burden) by increased resilience in situations that statistically occur e.g. once every 50 years can be politically challenging.

For reasons outlined above, the results of the modelling can be expected to be under more scrutiny than a typical peer-review process for an academic publication. This pressure is further increased by the fact that the scale of the problem necessitates limiting the granularity of the analysis from what would be possible on local or regional analysis. Thus, the modelling software, loss calculation process and all the related data management practices need to be monitored very carefully.

2.2 GAR contents

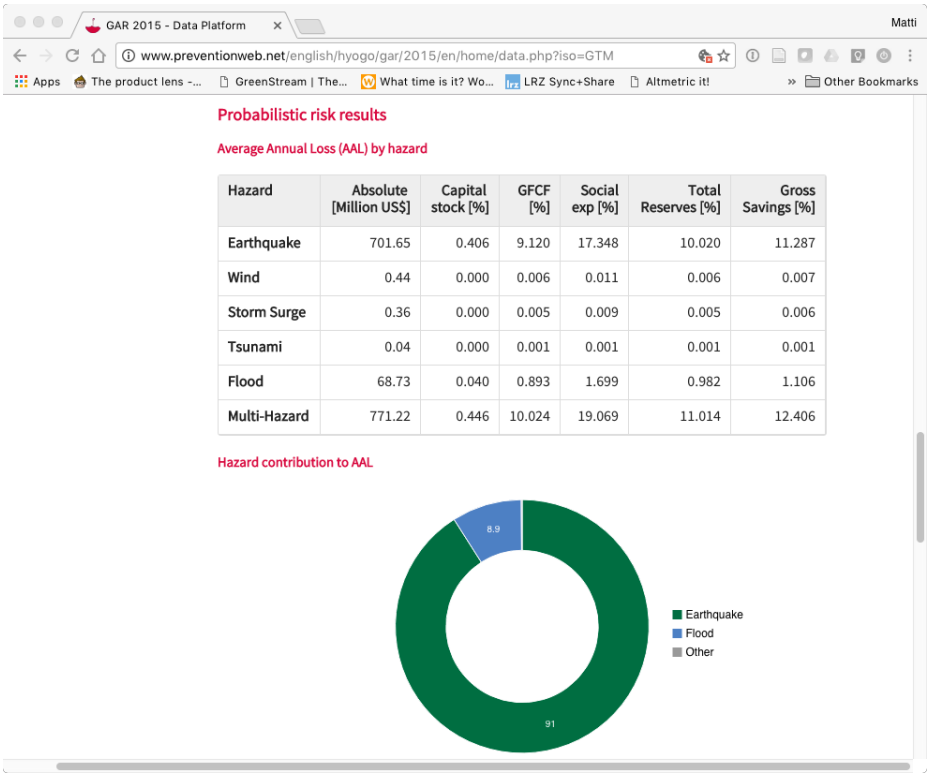
As outlined earlier, the GAR analysis process needs to take into an account changes in the hazards themselves (e.g. increased frequency of extreme weather events), the developments in the distribution of population and infrastructure (e.g. urbanisation) and the impact of the policies implemented so far (e.g. changes in building codes, flood barriers, early warning systems and so on). Based on input data from UN member states and from other sources (such as WMO and IPCC), a large-scale simulation effort will produce tens of thousands to millions of hazard-specific scenarios (e.g. flood, seismic, tropical storms etc.) describing the location and intensity of the natural hazard. The hazard- specific scenarios are put together with the country-specific exposure (geographical location and physical attributes) and their associated vulnerability corresponding to each hazard in order to perform probabilistic calculation of risk (likely losses).

The resulting risk metrics such as the AAL and PML for different return periods are put into context, e.g. by comparing the average direct economic losses to key macroeconomic indicators (see Figures 1 and 2). This can be used to gain a quick overview of the risks on the global scale (see Figure 3). An interactive viewer [10] is also available online.



Hazard	Absolute [Million US$]	Capital stock [%]	GFCF [%]	Social exp [%]	Total Reserves [%]	Gross Savings [%]
Earthquake	2,350.12	0.016	0.375	0.201	3.489	0.252
Flood	2,697.57	0.018	0.431	0.230	4.004	0.290
Multi-Hazard	5,047.69	0.033	0.806	0.431	7.493	0.542

Fig. 1. A small sample of some of the risk results of Germany as presented in the online version of the GAR. The average annual losses are put in the context of key macroeconomic indicators of the country, such as the Gross Fixed Capital Formation (GFCF) that measures the annual net increase of fixed assets owned by business sector, government and households and Social Expenditure. This contextualisation illustrates the load on the society natural disasters represent more effectively than mere average annual loss figures would.



Hazard	Absolute [Million US$]	Capital stock [%]	GFCF [%]	Social exp [%]	Total Reserves [%]	Gross Savings [%]
Earthquake	701.65	0.406	9.120	17.348	10.020	11.287
Wind	0.44	0.000	0.006	0.011	0.006	0.007
Storm Surge	0.36	0.000	0.005	0.009	0.005	0.006
Tsunami	0.04	0.000	0.001	0.001	0.001	0.001
Flood	68.73	0.040	0.893	1.699	0.982	1.106
Multi-Hazard	771.22	0.446	10.024	19.069	11.014	12.406

Fig. 2. Part of the Annual Average Loss data of Guatemala illustrate how the impact on society is considerably more severe despite the absolute losses being lower than the German ones (Figure 1)

The implementation of the Sendai framework for disaster risk reduction will increase the interest on both the GAR reports themselves, as well as the underlying data. The new framework for disaster risk reduction calls for more ambitious monitoring and modelling of risk. The plans to move into a “on demand” approach for risk modelling – as well as related policy developments that e.g. call for taking disaster risk into account when planning any investments – increase the demands on the assessment process. In addition, they make the provision of direct access to hazard, exposure and vulnerability data an important tool for increasing synergies between these activities.

Figure 3.5 Global multi-hazard average annual loss in relation to capital investment[11]

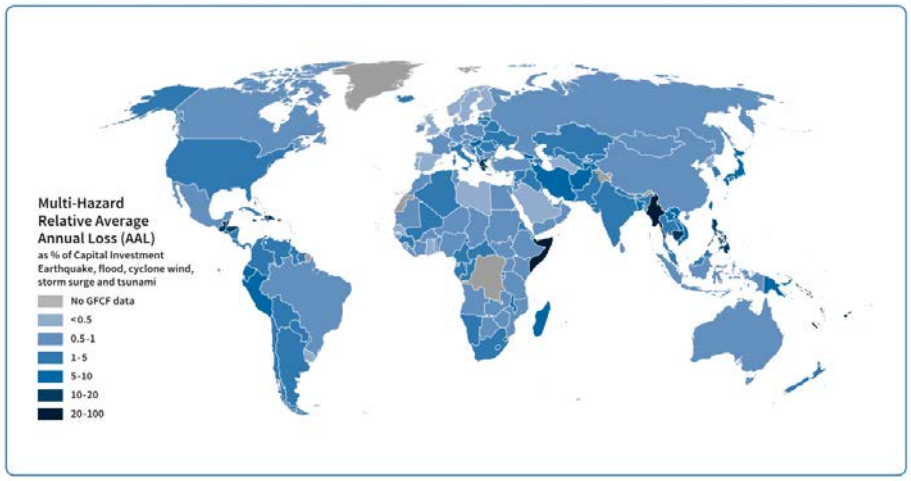


Fig. 3. A global summary of the proportional impact of the AAL to national economies

3 The GAR compute challenge

The GAR compute challenge can be broken into two parts: generating the hazard scenarios and combining them with the exposure database and the associated vulnerability functions to calculate the potential losses. The main dependencies between these steps are presented in Figure 4. The tools used in this process are heterogeneous, usually developed independently by the teams who are responsible for specific subtasks. Thus, the mode of operations is both globally distributed and very heterogeneous. On the technical level the CAPRA-GIS [11] toolkit plays an important integrating role: it provides the common formats for presenting hazard scenarios, as well as providing foundations for the overall loss calculation process. Hence the computational challenges have a clear interface between them.

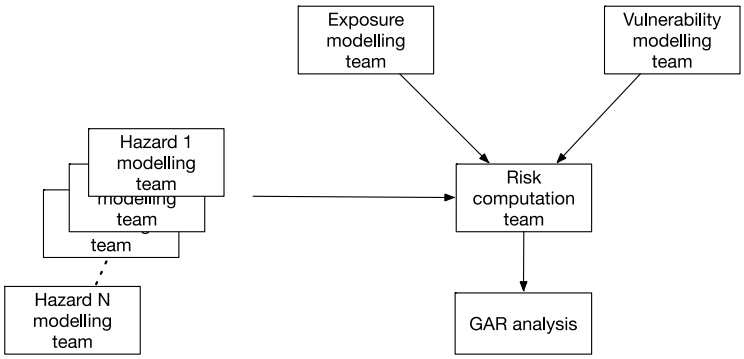


Fig. 4. Organisation of the teams involved in GAR analysis

3.1 Production of hazard scenario files

The hazard scenarios are produced using a wide variety of methods, ranging from models running on powerful desktop computers to ones using computer clusters to perform the work. The basic probabilistic process is similar: the hazards scenarios are produced independent from each other and can thus be considered trivially parallelisable processes. The implementation tools range from software developed completely in-house to models that run on platforms such as Matlab. However, despite the considerable computing resources needed, this step is rarely a time critical issue, as the schedule for producing the data is known well in advance and due to lack of dependencies between individual calculations any additional resources that can be brought in will speed up the process.

This situation may well change if the anticipated move to a more dynamic GAR process will be extended to the production of scenario files. For example, the overview document describing the latest approach to the production of the flood model data [12] lists 19 different external datasets used to initiate and fine-tune models, which represents a part of the process that is inherently serial in nature, representing an execution step that will not benefit from additional computing resources.

3.2 Loss calculation

The performance of the loss calculation step tends to be the most time critical part of the Global Risk Assessment process – already with the 2-year publication cycle. Any unanticipated delay in the generation of the hazard scenarios will delay the overall loss calculation. Furthermore, due to the importance of presenting the data on per-country

basis, summarising the loss estimates of large countries (such as China) can take a very long time.

Development of the CAPRA framework needs also to ensure that any new versions of the software produce same results as the original reference software versions. Hence, short-term developments tend to be incremental in nature. However, this incremental approach may face challenges with increased resolution (a uniform 1 km x 1 km grid instead of the current approach using resolutions ranging from 1 km x 1 km to 30 km x 30 km, depending on the hazard and geography of the region) and especially more dynamic production schedule of the GAR.

At the moment, the risk calculations are performed using two different versions of the CAPRA software: the original, single-threaded version (implemented in Visual Basic and available as part of the overall CAPRA-GIS package [11]) is used for most of the risks, as it is sufficiently powerful for risks where high-resolution modelling is not needed (e.g. modelling of droughts). A version of the risk calculation engine that has been parallelised and ported to Java (by the UNISDR team in collaboration with LMU and LRZ) is currently the reference implementation for flood modelling and is planned to be taken into use for the assessment of earthquakes in the near future. The Java version is already capable of exploiting shared memory parallelism and achieving close to a factor 50 speed up compared to single-threaded version. It is likely that further, incremental development of this version of the software will be sufficient to cope with the increase in resolution in the major publications (with the 2-year production cycle). However, additional optimisation is likely needed for the on-demand production of the reports (some potential approaches are discussed in chapter 6 of this paper).

4 The GAR data challenge

The current GAR data challenge can be broken down into two main phases: the hazard scenario stage and the data transfer to UNISDR. The vulnerability and exposure datasets are very small in comparison, and produced in more centralised manner. In either of the cases the amount of data is not a major problem per se, but rather the latencies introduced by the data production and transfer. While the hazard scenario development may in some cases need considerable resources, even the largest global hazard dataset will be of the order of few Terabytes. The situation might change slightly in the future due to higher resolutions used in the risk modelling. However, as the information is transmitted in compression format move from “5 km x 5 km” to “1 km x 1 km” resolution will not mean that the amount of data would automatically grow by factor of 25.

A bigger issue is moving the data to the loss calculation team in an efficient and coherent manner. While the current datasets are not excessively large (of the order of few Terabytes), limitations of the network infrastructure available to some of the partners in the collaboration necessitates moving the data by sending physical hard drives through postal or courier services. This approach is unlikely to cause insurmountable problems in the near future, as long as the resolution of the analysis will not be increased

dramatically. It is likely that the capacities of commodity hard drives will grow at sufficient rate to match the increase of the data volume.

However, the approach is not without its issues: copying the data from the original storage system to a transient media may introduce issues with consistency, especially if there will be additional versions of the data that complement the biannual GAR process. Developing processes and metadata approaches to handle these issues are relatively straightforward to manage in setting where the data processing is done by a relatively small, established collaboration. However, turning the hazard, vulnerability and exposure datasets into open data products and services used by a broader research community will likely bring up additional issues that GAR collaboration needs to address e.g. through additional documentation or training activities.

4.1 Impact of the on-demand process

Traditionally the GAR process has required storing only a complete, global dataset for each of the biannual publications. This means that even considerable increase in resolution will not increase storage capacity needed beyond what is possible to handle using commodity solutions. The two-year cycle will also create a framework for managing versions of the datasets in a very intuitive manner: even a directory structure that is based on the production year of GAR is sufficient in most cases.

The on-demand process will create additional challenges. When parts of the analysis will complement the main GAR dataset with updated information related to a country or a region, the consistency of the data management and the ability of the metadata system to maintain link between a publication and the corresponding dataset will need to be reassessed. With the increased resolution it is also possible that – if these additional model executions and subsequent versioning of the result datasets are a frequent occurrence – the size of the overall dataset grows to a level that makes moving it challenging. This may create a need to re-examine the current distributed computing model, as in more and more cases performing the computation in the same computing centre that holds the relevant datasets are stored will be advantageous or even necessary from the performance point of view.

5 Open data prototype and pilot

The motivation for investigating the feasibility of an open data approach are manifold, ranging from principles related to transparency of the GAR process to catalysing and supporting open innovation ecosystem that could also uncover novel approaches to disaster risk reduction. The data is already shared on request and used actively by third parties (e.g. by insurance companies to support their internal risk assessment processes). However, providing potential new users instant access is seen as a key method for removing barriers to new research and innovation activities.

The technical approach chosen for the open data pilot can be characterised as a "Minimal Viable Product" approach. The starting point is simple: a download portal addressing the key requirements of the GAR team and the external parties participating in the "beta phase" of the open data pilot:

- Download functionality: browse and download individual files or directories
- Mechanisms for branding and ensuring that users are aware of the licensing issues and key disclaimers related to interpretation of the data.
- Basic mechanisms for linking metadata to the data itself
- To be considered: upload/updates through the web interface – maintain consistency between different copies and with metadata

The first versions of the portal were developed at LMU in fall 2016, and will be used to refine the requirements of the production version that may eventually become part of the formal LRZ service portfolio.

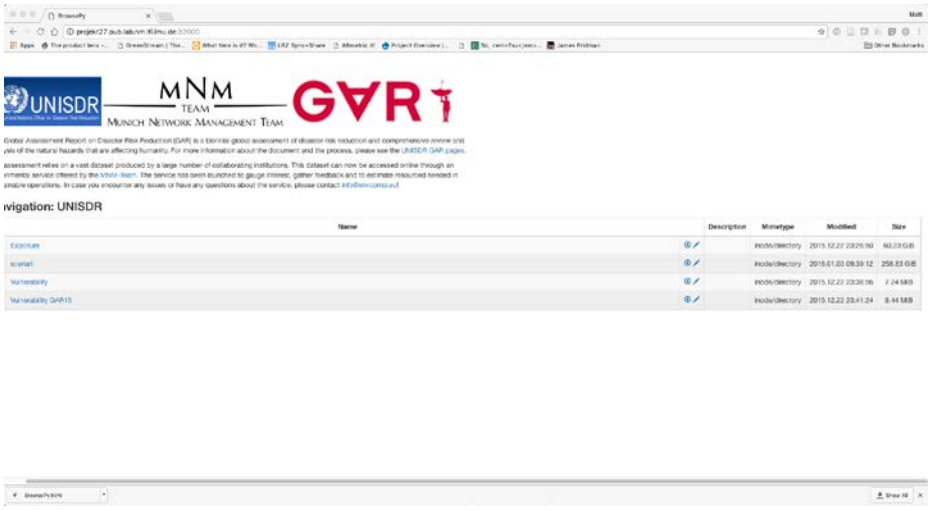


Fig. 5. A screenshot of an early version of the download portal (proof of concept) used to gather feedback related to the functionality.

Once sufficiently mature, the solution will most likely be merged with the current, download functionality [13] that is included in online version of the GAR, providing access to a limited subset of the information. The immediate needs for further development are mainly related to supporting archive file formats that are commonly used by the intended user community (possibly zip or rar formats in addition to tar.gz). In the longer term, aligning the data download functionality with the online visualisation tools and the Risk Atlas will also need to be considered.

Outside the purely technical issues, the open data pilot may bring up new requirements in terms of engaging with the community. Making the data available in a way that decouples it from the interpretations made by the experts (either in GAR or by the

groups who developed the hazard models) may bring up situations where the interpretation – or misinterpretation – of the data outside the strict UNISDR scope should at least be brought to the attention of the original experts. Thus, supporting functionality such as social media integration may be deemed a useful complement to the download functionality itself.

6 Future directions

The open data portal will most likely play a role in the launch of the 2017 version of the Risk Atlas that provides a consistent geographical view of the disaster risk information, allowing business, investors and international organisations assess and compare the risks and resilience to the occurrence of natural hazard events in different countries more intuitively. Linking the GAR data organically to such an overview document will most likely increase interest in both. Thus, determining methods to discover and support new use cases emerging from the use of Risk Atlas and GAR data will proceed in parallel with the implementation of the first version of the open data service.

We foresee that the data management back-end will need additional functionality as the dynamic, on-demand approach to the production of GAR and country reports will be adapted. In the absence of a clear two-year cycle of report production, issues such as persistent identifiers, versioning of the data and metadata need to be reviewed. Initial assessment of solutions such as KIT Datamanger [14] and CKAN [15] as tools to meet these new requirements is already ongoing.

In response to the computing challenges related to risk calculation outlined in the chapter 4.1, more advanced parallelisation approaches are already being studied by the collaboration. For example, both MapReduce [16] and MPI [17] based approaches could allow parallelisation beyond the limitations of the shared memory space and – with sufficient resources – even allow in-memory processing of the data.

The new data management approaches and tools discussed will obviously also have an impact on the processes that are internal to GAR collaboration manages the data. They may also influence the interfaces the GAR data services can support for the third-party analysis. As an example of a potential explorative research topic, investigating approaches where the analysis of the data could be performed at the storage location in a flexible manner (e.g. by shipping the analysis code embedded in a virtual machine or software container to data) could support more efficient and flexible management and use of growing and increasingly dynamic GAR data.

7 References

1. UNISDR: Global Assessment Report on Disaster Risk Reduction (GAR) 2015, http://www.preventionweb.net/english/hyogo/gar/2015/en/home/index.html
2. The CERN Engineering and Equipment Data Management Service, https://espace.cern.ch/edms-services/default.aspx
3. Worldwide LHC Computing Grid, http://wlcg.web.cern.ch/

4. GEO-DARMA = Data Access for Risk Management, Group on Earth Observations, https://www.earthobservations.org/activity.php?id=49
5. UN-ISDR: Hyogo Framework for Action 2005–2015: Building Resilience of Nations and Communities to Disasters. United Nations – International Strategy for Disaster Reduction, UN/ISDR-07-2007
6. United Nations General Assembly: Sendai Framework for Disaster Risk Reduction 2015 – 2030, A/CONF.224/L.2
7. Intergovernmental Panel on Climate Change, https://www.ipcc.ch/
8. OpenStreetMap Foundation, OpenStreetmap Project, https://www.openstreetmap.org/
9. UK National Audit Office report, "Strategic Flood Risk Management", p.38, paragraph 2.26, ISBN: 9781904219460
10. UNISDR, Global Assessment Report on Disaster Risk Reduction 2015, Risk Data Platform CAPRAViewer, http://risk.preventionweb.net/capraviewer/
11. CAPRA Probabilistic Risk Assessment Program, CAPRA-GIS, http://www.ecapra.org/capra-gis
12. Rudari R. et al, Improvement of the Global Flood Model for the GAR 2015, Input Paper prepared for the Global Assessment Report on Disaster Risk Reduction 2015, http://www.preventionweb.net/english/hyogo/gar/2015/en/bgdocs/risk-section/CIMA Foundation, Improvement of the Global Flood Model for the GAR15.pdf
13. UNISDR, Risk Data and Software Download Facility, http://risk.preventionweb.net/capra-viewer/download.jsp
14. WM, KIT – Universität des Landes Baden-Württemberg and nationales Forschungszentrum in der Helmholtz-Gemeinschaft: Kit Data Manager – The Research Data Repository Platform, http://datamanager.kit.edu/
15. CKAN Association, CKAN – The open source data portal software, http://ckan.org/
16. Dean, J. and Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters, OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004, https://research.google.com/archive/mapreduce.html
17. MPI Forum, MPI Documents, http://mpi-forum.org/docs/