Eawag directive on the archival of research data

1 Preamble

Eawag recognizes the fundamental importance of research data in maintaining quality research and scientific integrity, and is committed to pursuing the highest standards. data management standards [we should say highest standards regarding what, I suggest "data management", an alternative could be "research integrity"] Eawag acknowledges that correctly and easily retrievable research data are the foundation of and integral to most researchs projects. They are necessary for the verification and defence of most research processes and results. Research data have a long-term value for research and academia, with the potential for widespread use in society.

This directive is based on and concretizes the regulations regarding the collection, documentation and archival of research data set forth in the Guidelines for Good Scientific Practice[1].

2 Terms and definitions

2.1 Research data

Research data refers to all information which is commonly accepted in the research community as necessary to validate research findings, including contextual information. *Research data* include all materials which are created or utilized in the course of academic work, including digitisation, records, experiments, measurements, observations and surveys and interviews. This includes software and code. *Research data* can take on several forms during the lifespan of a research project and includes gradations of raw data, processed data and published data.

This directive only refers to *research data* in digital form. Other incarnations of *research data* need to be dealt with on an individual basis, although the general principles laid out here apply.

2.2 Research data of archival value

Research data of archival value are all *research data* that meet at least one of the following three conditions:

1. The data are necessary to reproduce or replicate the results of an *Eawag publication.*
2. The data are costly to reproduce or not reproducible at all (e.g. all observations and measurements in the environment).

   Maybe this is a bit too strict. Should we leave some room for decision to the departments? E.g. data of considerable size and quality or similar? Otherwise, each minor data sampling e.g. with students would have to be archived.
3. The data have to be preserved in the long-term for legal reasons.

[1] Eawag Directorate (2014). Research Integrity at Eawag: Guidelines for Good Scientific Practice. http://www.eawag.ch/fileadmin/Domain1/About/Arbeiten/Forschungsumfeld/integritaet_forschung.pdf.

Hier sollte noch so etwas kommen wie "2.2.1 Maturity for archival" das sind die Daten "of archival value", deren Auswertung abgeschlossen ist und die "statisch" sind. Die kommen auf ERIC. Die "pre-mature" auf den zentralen Storage.

Die wichtigste Bemerkung ist eigentlich der Vorschlag, diese Directive und deine Interne Eawag-IT Policy: Migration grosser Datenmengen auf die zentrale Infrastruktur in einer Policy zu vereinen, die dann einfach Eawag directive on the storage and archival of research data heissen wuerde.

2.3 Eawag publication

An *Eawag publication* is a scientific publication (including peer-reviewed articles, book chapters, thesis and gray literature) that has at least one Eawag-affiliated co-author.

2.4 Eawag Research Data Institutional Collection (ERIC)

ERIC (https://data.eawag.ch) is an Eawag internal web-based repository for research data. It is run and maintained by the IT department and provides a convenient way for the archival of research data in conformity with this directive. The primary entity of storage in ERIC is the *data package*.

A publicly (outside Eawag) available instance of ERIC contains the meta-data of all packages and also the actual data of packages that have been flagged as *open data* by the responsible PI or Group Leader.

2.5 Data package

A *data package* is comprised of an interrelated set of *research data* that is quality-checked, organized, documented and formatted for archival. The data in a *data package* is documented to a standard such that they can be understood and used in a scientific context by a researcher who was not involved in the original research and has no access to members of the original research team. In general, a *data package* is stored in ERIC.

Exceptions: ? etwa Branchen-uebliche repositories (Genetik etc.)

2.6 Publication data package

A *publication data package* is a *data package* that contains all data and ancillary information necessary to reproduce or replicate the results of an *Eawag publication*.

I suggest to be a bit less strict in documentation requirements for publication data packages, because these should contain scripts etc. for all data processing steps and may not be reproducible in 20 years due to lacking availability of compatible software. Unique data underlying the publication still has to be archived as a data package according to the guidelines. But this package does not have to contain all processing scripts for the publication.

2.7 Open data

Open data are *data packages* that are made available to the general public for unrestricted download, preferentially through the public facing instance of ERIC. A persistent identifier (DOI) is assigned to each *open data* package. *Open data* is being made available under appropriate legal terms to maximize reuse and dissemination.

Unless prevented by issues regarding competitive advantages, the protection of personal data -> personally identifiable information, intellectual property issues -> rights brought about by external collaborators or other legal problems, Eawag strongly recommends to publish all research data as *open data*. This increases the visibility of the respective research group, of Eawag in general, increases the probability to be cited and to establish future collaborative projects and optimizes the value of Eawag's output for the scientific community and society at large.

3 Responsibilities

3.1 Group Leaders and Principal Investigators (PIs)

In conformance with the Guidelines for Good Scientific Practice, the PI of a research project is ultimately responsible for the proper annotation, documentation, organization, formatting and reliable long-term storage of all *research data of archival value* that is produced in the course of the project.

ergänzen: He/she is the data owner.

In particular, **the PI is responsible that for each *Eawag publication*, a *publication data package* is uploaded to ERIC**. The *publication data package* should ideally be deposited in ERIC immediately after the associated publication has been accepted. Jochen: -> "after the associated publication has been published"

Should we limit this to publications with first or senior author from Eawag? It seems unrealistic that we can do this for all publications to which we contribute, but the key responsibility with with an external institute. DORA does not know which authors belong to Eawag ??

Exception: *Eawag publications* with external collaborators, for which either the associated data itself or the necessary expertise to annotate them is not available at Eawag, are exempt from this rule.

The Group Leader takes care that new workgroup members are made aware of this directive and organize their work accordingly.

The Group Leader identifies *research data of archival value* which is not deposited as part of a *publication data package* and takes care that this data is packaged and stored appropriately.

It is recommended that Group Leaders establish workflows and protocols that lead to an efficient data management throughout the research lifecycle.

3.2 IT Department

The IT department

a) runs and maintains the Eawag-internal research data repository (ERIC), which provides a convenient way to annotate and store research data in compliance with this directive.

b) plans for and acquires the necessary storage capacity.

c) provides up-to-date infrastructure and interoperability with external services to widely disseminate *open data*.

d) provides guidance and support regarding research data annotation, preparation and deposition in ERIC.

e) provides guidance and support regarding the publication of software in the context of *open data*. You do not say anything about data formats. In my view, the IT department should be responsible for converting data to relevant formats for the long-term archive (not for publication data packages). This requires storage in a format accepted by the IT department for this purpose (e.g. pictures, videos, etc.). For publication packages, this cannot be required as easy reproduceability requires also the acceptance of proprietary formats.

3.3 Lib4RI

Can Lib4RI contribute anything here? Help with quality checks of data packages for example.

Lib4RI hosts the associated publications in DORA.

3.4 Directorate

The directorate provides the necessary personal and financial resources to ensure the long-term operation of ERIC, the acquisition of the associated storage capacity, and the associated support activities.

4 Validity

This directive is subject to review one year after coming into effect.

2.3: Definition Eawag publication:

Is peer review a requirement for all relevant publications or only for journal articles? I would recommend to restrict this definition to peer-reviewed publications. This would reduce the relevant publications more or less to journal articles. Besides journal articles there are not many Eawag publications with peer review in DORA. For other publications than journal articles it is difficult to assess if the publication is peer reviewed or not (for the library staff). Thus the peer review status for many of these publications does not exist in DORA (or is not reliable).

The definition of an "Eawag publication" given here is important if one wants to check if for all relevant publication in DORA have an associated data set in ERIC. Under the current definition more or less all publications in DORA should have a dataset deposited. This will produce many false positives for publications other than journal articles.

A thesis never has an Eawag-affiliated co-author. The affiliation of the author is always the degree granting institution. Eawag supervisors are not authors of these publications.

If theses should be included in this definition I would recommend to limit this to dissertations (and not master or bachelor theses). Because there is quite often some confusion about "peer review" here is a defintion: "Peer review is the system used to assess the quality of a manuscript before it is published. Independent researchers in the relevant research area assess submitted manuscripts for originality, validity and significance to help editors determine whether a manuscript should be published in their be published in their journal.journal.