Towards a standardised metadata system for physics and related fields of research and development

Context: Metadata descriptions and schemas are the cornerstone of good practices in research data management (RDM), as envisioned by the FAIR principles. Metadata are the 'languages' used by measurement instruments, data analysis and validation software, data archives and repositories, and finally exploited by data search engines. In order to make data interoperable and re-usable for the sake of the scientific community and society as a whole, these 'languages' must be harmonised in a way which is scientifically founded, adherent to best practices and comprehensible for humans and machines. The consortium NFDI4Phys is going to submit a proposal to Deutsche Forschungsgemeinschaft (DFG) in the 2020 funding call, to participate in shaping the RDM landscape for physics in Germany, guided by the FAIR principles.

The state of metadata systems and elements in physics: The vast majority of research data in physics and other quantitative natural and engineering sciences can be, simplifying, characterised as *numerical factual data*: each individual data point represents a quantity, consisting of a numerical value, the appropriate unit and an estimate of its uncertainty. Descriptions of the research object and the generation (and transformation) methods provide necessary context for their correct interpretation. However, no existing metadata standard addresses numerical factual data comprehensively. Various research fields, groups, or individual scientists have been designing descriptions to grapple with their specific needs. A whole 'zoo' of sometimes similar, yet often incompatible schemas provide solutions tailored for specific types of experiments. Large-scale experimental collaborations have had an advantage when it comes to agreeing on quasi-standards for their respective communities. However, for RDM outside these use cases, solutions are lacking. The current 'zoo' of metadata systems contradicts the stated goal of interoperability: because the number of necessary converters rises very quickly with growing numbers of concurrent quasi-standards, increasing, too, the vulnerability towards interpretation errors and the maintenance costs of converting tools.

NFDI4Phys: The NFDI4Phys consortium aims at addressing these needs by bringing together a broad group of stakeholders to bridge the gaps between the existing isolated metadata systems. Drawing from the present methods and working with researchers from all interested fields in physics, we intend, on the short term, to develop prototypes that can supplement the current systems and, on the long term, overcome their limitations. Through foundational advances and by suggesting and advocating ambitious quality criteria and best practices (that are going to evolve with time), we are planning to foster the emergence of long-term sustainable community-wide metadata standards.

Our approach: NFDI4Phys is going to focus as a priority on the theoretical foundations (ontologies, thesauri, metadata modules, application guidelines) of the future metadata standards as well as their hands-on implementation into flexible, easily usable, modular tools. Our guiding principle is depicted in the viewgraph: diverse descriptions are going to be united in a moderated process aiming at their subsequent alignment with emerging, continually improving standards. *Pilot users* selected from a wide range of research areas will drive the development. Easy usability will be crucial; the modular design allows for a continuous upgrade cycle. The ideal end state is characterised by a small number of inter-related and federated real metadata system standards which, taken together, cover the metadata needs of not only the physics community, but also other fields of research and development driven by numerical factual data. These standards should meet high quality specifications and be interoperable also in an international context, i.e. they should be portable to different natural languages and implementable using different programming and markup-languages and description syntaxes.

Collaborations outside the NFDI: The establishment and continuing improvement and maintenance of the SI (International System of Units) offers a successful historic precedent as to how the today's RDM challenges can be met via international harmonisation. PTB as the applicant institution for NFDI4Phys is able to provide the experience and trustful relationships with its international counterparts to foster the necessary international coordination for establishing harmonised metadata standards. Moreover, the consortium intends to liaise with standardisation bodies and manufacturers of measurement instruments towards the implementation and adoption of the future metadata system standards into digital devices (hard- and software) coping with numerical factual data.

Conclusions: Committed to the FAIR principles, the NFDI4Phys consortium promotes FAIR research data management within the physics community in Germany. Besides metadata standards, we are going to address other relevant issues, such as data quality metrics and data repositories. Following a bottom-up approach, the development priorities will be based on the needs and demands expressed by the community via continuous feedback, starting with an ongoing survey. NFDI4Phys stands out in the breadth of the potential community and in the fundamental approach based on the shared metrological concept of numerical factual data. Our main target audience are physicists whose RDM demands are not covered yet; our initiative thus intends to complement the other emerging consortia from disciplines of physics. Yet the fulfilment of the FAIR principles can be achieved only by joint efforts by the scientific community overcoming the subject borders. Therefore, NFDI4Phys will cooperate strongly on cross-cutting topics with other consortia from physics and neighbouring sciences, in order to ensure harmonisation and compatibility of standards and procedures among all quantitative scientific disciplines.

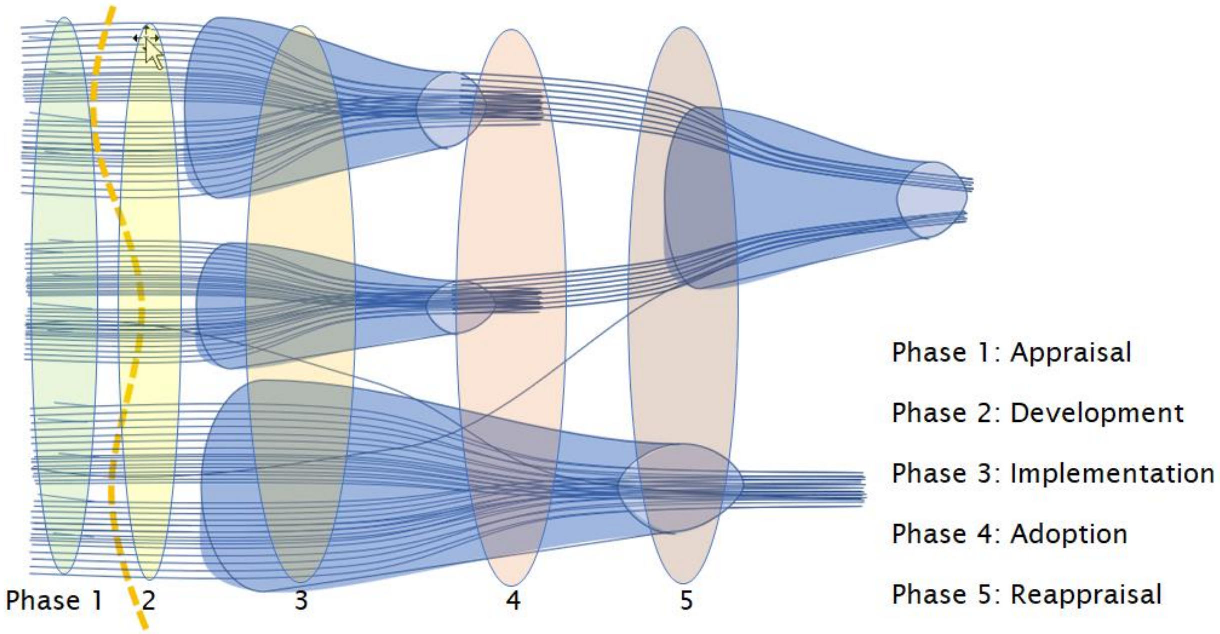

Figure: The wide variety of metadata quasi-standards for numerical factual data gains cohesion due to the standards that are going to be introduced during the NFDI process. The blue envelopes represent examples of the small number of metadata (proto-)standards emerging from the NFDI process. After the appraisal of the metadata landscape in phase 1, proto-stadards are being developed in phase 2, and implemented in phase 3. In Phase 4, users gradually adopt the new standards instead of older formats. The reappraisal in phase 5 completes the first iteration of the process, which can consists of several cycles.