GO FAIR

IN definition

MEMO GO FAIR

6 July 2018 at 07:07

Topic:	**FAIRwizard**
Interest:	making FAIR easy
Distribution:	GFISCO, Mark Musen, Robert Pergl, Rob Hooft, Annalisa Montesanti, Margreet Bloemers, Peter Wittenburg, George Strawn
Author(s)	**Barend Mons, Erik Schultes, Luiz Bonino**
Main goals:	To develop, deploy and offer a lay-person proof, comprehensive DSPlanning environment for data, tools and compute resources, that can also be operated by funders to specify DS requirements
Reviewers:	candidate founders
For info to:	HRB, ZonMW offices, Belmont Forum, ORFG, George Strawn, NIH, NSF, NIST etc.

The background documents about GO FAIR can be found at the GO FAIR Website
The report of the European Open Science Cloud can be found here

Preamble

It is assumed that the readers of his GO FAIR memo are sufficiently aware of the basics of GO FAIR and the FAIR principles it is based on.

The *de novo* creation of research outputs as digital objects as well as their proper management and stewardship over longer periods of time is a crucial part of modern, data driven search environments and infrastructures. More often than not, nowadays we will re-use Other People's Existing Data and Services (OPEDAS) for our own research. This puts the question of *reusability* of our research outputs center stage for good data stewardship.

So, researchers, especially those receiving public funding for their research, should consider the societal responsibility to consider reusability of their research outputs and to ensure that their research outputs are Findable, Accessible, Interoperable and actually Reusable (FAIR) as long as useful and possible. As specified in the FAIR principles, reusability goes beyond mere findability, accessibility and interoperability (by humans and machines). Data that are perfectly F, A, and I can still be (re)useless, because they are of low quality and the metadata and other contactual information are insufficient for people, and specifically machines to decide on reuse for what purpose and in which work flow environments. Next to that the actual 'quality' of dat can be bad. In essence even completely fabricated data can be perfectly FAI, but should not be re-used (other than for detecting plagiarism, fraude and pseudo-science). In any case, rich metadata, carefully describing the scope of the original research, the methods used, and rich provenance throughout the entire research process are always critical for third parties (again, machines and humans) to decide if, and for what research, objects (both data and related tooling and services) are practically reusable for their purpose, which may in many cases be different from the exact purpose for which these resource were originally created. This also means that proper metadata capture and publishing is at the core of modern data stewardship and allegedly among the most challenging parts of good data stewardship plans.

The issue:

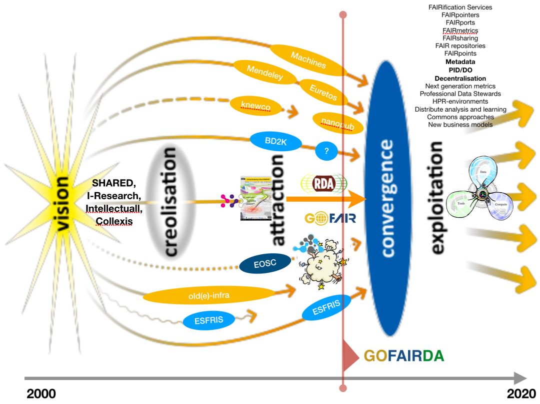
Apart from the enormous benefits for research of having a growing Internet of FAIR data and services, research funders also increasingly request and monitor good datastewardship planning and execution as part and parcel of each research project they fund. In the public funding sector, the emerging trend is now to request (planning for) FAIR data and sufficiently rich FAIR metadata for datasets, papers, software and other research outputs. The trend is also to request 'open data' wherever possible and only allow opt-out for full open access in case a valid argument can be made for restricted acces, such as personal privacy or national security reasons. Even in these opt-out cases, the data as well as the metadata should still be made FAIR (which is not equivalent to Open) and the metadata could be made open, independently from the decision of the level of accessibility of the data itself.

The emerging issue:

It is one thing for funders to endorse the FAIR principles and to request adherence to these principles in data management and stewardship plans, but is is a *whole different issue* to consider how much time and effort is requested from already overburdened researchers with little data management skills, beyond their own use of the data they create. It is only 'fair' to ask a level of FAIRness in each particular research project or research Call/RFA that is achievable with reasonable effort.

Very importantly, preparation, construction, evaluation and certification of good data stewardship plans, as well as the inherent costs of their execution after approval of the proposed projects should be considered part and parcel of modern research and they should be eligible/allowable costs to budget for in the proposal. The ballpark and average (g)estimated percentage expected to be involved in the entire process of planning for, construction and execution of DS plans is 5% of the research budget, whereby obviously a wide variation is expected (examples of 1% as well as close to 50% have already been recorded).
It should be noted, that the 'perceived market' for tools in the DS realm is than estimated at about 100 billion annually world-wide, and at 10 B for European, publicly funded research only. With an estimated loss[1] of well over 10 B annually in the same realm for Europe alone, the realisation starts to settle in that the investments in good datastewardship would have a very favourable RoI profile and in fact 'pay for themselves' form days one. This also attracts major and classical private providers of research support tooling to follow FAIR principles in order to ensure a part of this rapidly developing 'market'. As we have seen several disasters in the past when core services for the scientific community were left to uncontrolled monopoly developments, we need to be fully prepared to prevent such monopolies, which appear to be bed for everyone in the end (including for the now deeply resented monopolists). This is why GO FAIR seeks PPP's from the onset with clear rules of engagement including monopoly and vendor lock-in prevention.

Next to considering the costs associated with the entire DS cycle as eligible for funding, funders have now shown increasing interest in providing researchers (as well as professional data stewards) with **user friendly tools for the construction of proper datastewardship plans**. There are also GO FAIR, EOSC/Commons, RDA and IMI related projects aimed at creation of the actual tooling to make existing and *de novo* research objects FAIR in practice as well as metrics to measure the actual levels of FAIRness.



What is the current state of attraction & convergence?

In short, many forces have come together to ensure commitments to FAIR Principles (often very publicly) from funding agencies, universities, and industry. But this is, so far, very *aspirational* (its a bit like "world peace"… who can be against?). So we see there is already demonstrable need for FAIR solutions. However, the implementation of FAIR in practice is still, for most people and organisations, somewhat of a mystery and maybe in the state of late creolisation in the Strawn-Wittenburg sense, see picture and reference-to be added).

This tension creates a state of 'super saturated'

[1] PWC report when out

IN definition

FAIRness in the community (Europe as well as USA).

How to move from attraction to convergence?

We have recently discovered (in a series of GO FAIR workshops and hackathons) a way to arrange a few key pieces of existing technology to trigger crystallisation in the FAIR Ecosystem (accelerate attraction and hopefully later convergence in the Strawn-Wittenburg sense).

These key technology components are:

(1)**FAIR Metrics** https://www.nature.com/articles/sdata2018118
(2) Data Stewardship **Planning Wizard** https://dsw.fairdata.solutions
with Open Access links to a **DS book** https://www.taylorfrancis.com/books/9781498753180
(3) **CEDAR** https://metadatacenter.org

A combination of DS planning tools, metrics and metadata construction environments could bring the breakthrough, as experienced in the recent Wizard+Metrics hackathon.

As a side-effect of that hackathon we came to realise that there are, in practice, only 3 data stewardship planning tools that have wide aspirations and already some community uptake: **DMPOnline** (DCC), **DMPTool** (CDL), and the **Wizard** (Elixir-CZ/NL, DTL). These tools are in fact quite different, to a large extent complementary and each their strengths and weaknesses. Interestingly, it turns out that DMPOnline and DMPTool are already in the process of merging.

We also see (through BioPortal and FAIRsharing) a more and more complete picture of existing ontologies, templates and best practices for many 'semantic types' that appear in data and metadata across disciplines. In addition, specific efforts are underway (a.o. in the scope of EOCS) to further define metadata templates. These semantic types include concept categories that are important for the 'scholarly record' such as people, institutions etc. (see for instance **VIVO**, also aligning with GO FAIR and FAIR principles).

How can GO FAIR bring this all together?

1. **GO FAIR will organise a workshop (1)** focused on the top 3 tools in order to align future development in a coordinated matter, and bringing the DMP and DSP tool developers together. The mission would be to simplify the DMP and DSP landscape for the users (funding agencies and data stewards), creating a single, complete, and trusted tooling package [see broad design later on]. It is expected that Workshop 1 will result in a manifesto for a new GO FAIR IN called the FAIRwizard IN. Here is a draft manifesto: https://docs.google.com/document/d/1Z37a1humM0gY6oKQWrxquHOYid-qcJxnGK2nFoJefDA/edit?usp=sharing

2. **Integrate metrics in DS planning tools**: There are for now 14 automated metrics, that together require 22 inputs (all spelled out in the paper and also here: https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf).
However, many of these inputs require choices for data standards, ontologies and the creation of metadata templates. Although many of these choices and templates will have generic and overlapping components, they must nonetheless be defined by each stakeholder community that wants to use the metrics. To make things easier for stakeholders, We have made these community decisions explicit as a set of 29 "Community Challenges" (appended below… but also now in a manuscript currently reviewed at Nature). Essentially, these 29 Community Challenges frame the decisions that need to happen in order to make FAIR increasingly standardised and automatic. They also frame the tension between importance of the FAIR Principles, and the 'cost of compliance' (and so help to frame budgets in project proposals).

The Community Challenges ask [ultimately] for **domain specific metadata schemata**, standards and protocols that enable machine-actionability regarding F, A, I, and R. So helping FAIR-minded communities to define and then implement metadata profiles and metadata capture is an obvious and natural application for the FAIRwizard, including major projects such as CEDAR for automated, user friendly and FAIR compliant metadata templates. Of course, work in metadata is already highly advanced, but is still in the 'creolization phase' [reference] and has yet to really 'converge'.

3. So, we also conceived **GO FAIR/RDA (sub)Workshop 2**: Metadata for Machines. The idea here is to assemble (again 1.5 days, in Leiden) the 'critical mass' in the field of metadata, along with CEDAR and others [maybe the Peter Doorn/NWO led initiative?. The goal is to define the metadata requirements for the FAIRwizard IN and to hammer out the minimal machine-actionable metadata (atomic templates, see below) that are required inputs to metadata templates and to answer the FAIR metrics (taking a lead for the 29 Community Challenges). The generic solutions achieved in this group will be offered to others as "GO FAIR" community emerging solutions and made freely available, in an attempt to drive 'convergence'. Peter Wittenburg and Erik Schultes are now sketching out the parameters for this workshop: https://docs.google.com/document/d/1LfnPnTca0WTkMc7s8sSem-Qv1BuTcGFrZGaNVklQKJQ/edit?usp=sharing

The hypothesis is this: Of the many ways to 'nucleate' the crystallisation of a FAIR ecosystem, automating FAIR resources in the combined **FAIRwizard** concept is the most efficient. Why? Because many different stakeholders are likely to take their cue from Data Stewardship Planning tools, especially if funding agencies require it.

Although hypothesis, we think we already see evidence that this is may be true. First, the already-mentioned interest from ZonMW and HRB. Second, last week a leading GO FAIR Implementation Network, the Chemistry IN, held a two-day meeting to define their community (IUPAC) specific data standards and FAIR metadata templates: https://www.eventbrite.co.uk/e/fair-chemical-data-workshop-2018-tickets-42482231498. Although there was, as one would imagine, some struggle in the ChIN to get a grip on the big issues, here is the schema that emerged.

Again, the point behind all this "support and coordination' effort in GO FAIR is to accelerate the Strawn-Wittenburg convergence. Technology alone is never the solution, but given the political and economic context, it may be that bottom-up solutions could trigger a field supersaturated with FAIR potential.

Our next move in GO FAIR is to organise Workshops 1 & 2 as back-to-back events in Leiden (August or September). Clearly CEDAR has a key role to play in this configuration as does development and training.

Focus of the FAIRwizard IN-spe/Project:

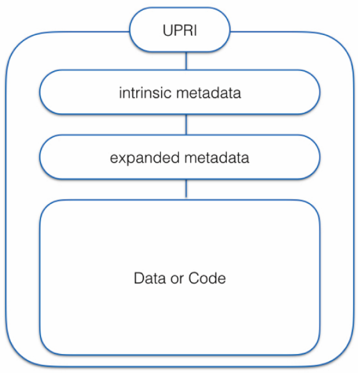
Tooling for the actual planning and construction of data management and stewardship plans is still in very early stages. We will bring together the major groups around the globe that have parts of the required suite of solutions, best practices and actual tooling and we have the ambition to collectively develop an end-to-end solution/workflow for data stewardship planning that is fully aligned with developing best practices in other Implementation networks (and beyond) in oder to make it as easy as currently possible for funders to request appropriate and reasonable data stewardship and FAIRness level, and as convenient and straightforward as possible for the applicants to comply with the requirements and develop proper data stewardship plans that fulfil the criteria of the Call/RFA.
Obviously, data stewardship requests may vary significantly between funders, and also per discipline. So we need an environment that caters for a wide variety of 'levels of FAIRness' of research outputs and that can be easily adapted for each Call/RFA as well as scales with the increasing possibilities to make data FAIR at higher and higher levels. An important point in Europe is -for example- that there should be a minimal-required DS FAIRness level that can be adopted by the EC funding programmes with an equal opportunity for participation of all eligible countries, whilst allowing easy additional retirement setting for national funding schemes in countries that are spearheading data stewardship.

Trying to focus on the most important things first:

There are relatively simple and high level decisions to be made in a data stewardship plan that do not need specific tooling or deep knowledge about data modelling, metadata schema's, ontologies etc. These include for instance: where to publish papers (in Open Access or certified journals, with a guaranteed permalink to the data or software they describe), where to put the data (in a trusted and certified repository), how long is storage of the data required by the funder or the institution, and who will cover the associated costs. What license is given to the data and so forth. These questions, **although critically important for data stewardship planning** do hardly need novel tooling, although researchers should be assisted where possible in finding machine readable solutions, making decisions at these points and answering to funders or publishers requirements.

However, we decided that such generic questions are largely outside the scope of this IN and will be answered in developing handbooks, other INs, such as OPEDAS and GO TRAIN.

The FAIRwizard IN will focus mainly on one of the most 'technical' aspects of data stewardship that is also an area where many researchers are not trained for, and which is nevertheless a critical prerequisite for data and services to participate in distributed systems for reuse of data and tooling for research. The actual routing of data to tools, tools to data and both to come together at a place where there is sufficient compute to run the job at hand, all depends on the richness and quality of the machine interpretable ***metadata*** associated with these three major elements of the IFDS.



The **C2CAMP IN** is dealing with the actual routing infrastructure, but the 'directory' for efficient routing is in FAIR, machine actionable metadata for all elements of a study. Capture and publishing of FAIR metadata is therefore **a central requirement for good datastewardship** but most project officers, researchers and publishers (to name a few stakeholders) are not yet proficient in the technicalities associated with FAIR metadata. These include for instance deep understanding of conceptual modelling, linked data formats, ontologies, protocols and the crucial difference between 'intrinsic' and 'user defined' or 'expanded' metadata (effectively annotations pertaining to the research object the metadata are 'about'. (see figure 1).

Rich metadata are also needed to track the use and reuse of data and tooling to cite the and to link them to scholarly collections per person, per institution and at higher levels of aggregation. With the rapidly increasing frequency of major analytical processes involving multiple, distributed datasets and tooling, it is impossible to track reuse and proper citation of all these elements in a manual way. So, also here, machine readable scholarly information at the personal and the institutional level, interoperable with analytical frameworks, data and text publication environments, citation and attribution systems will be key (hence the involvement of VIVO in this IN).

The intricacies of how this can all be achieved are described in reference 1, including a block-chain type tracking approach, but here we will only describe the very basics and explain why the networks of excellence that join forces in this IN are all complementary and needed and how they represent a critical mass for broad community impact.

The **'funders' aspect**, representing early movers in the requirement for FAIR outputs and recognising the responsibility to make these requests precise, practical and achievable for researchers is represented by [x] funders [names, ZonMW, HRB, NIH, NSF etc.]. These are mainly from the life sciences and health field, but we argue that privacy and complexity issues around data are very challenging in this particular domain, so we take on the 'hardest challenge'.

The **Metadata aspect** and the ontology/standards aspect are represented by [CEDAR, Bioportal and FAIRsharing, as well as by the conceptual modelling group [Luiz etc.]

The **metadata and data publishing** aspect is covered by the technical development group [led by Luiz] that has already implemented basic reference technology such as FAIRpoints [] and FAIRports[] based on open standards and open API's to publish metadata.

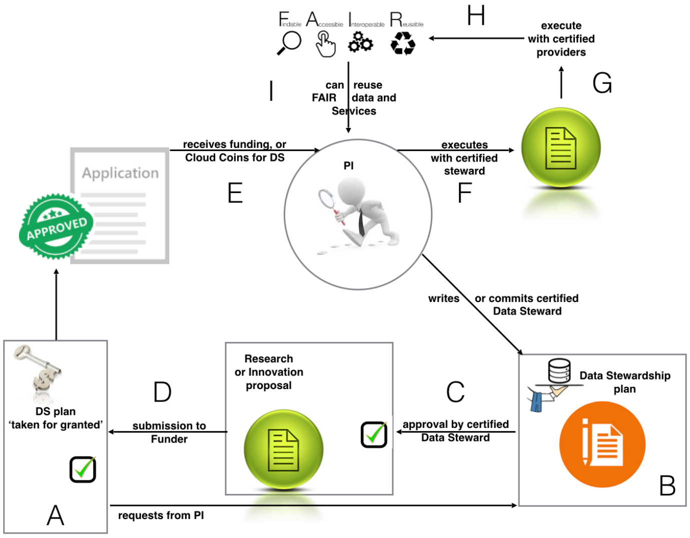
The **scholarly connection** and attribution aspect is represented by [VIVO, others?]

The **DMP tooling aspect** is represented by DCC (DMP online, the Data Stewardship Wizard development team [co-lead by Robert Pergl and Rob Hooft, Czech republic, Netherlands) and …..[please add], who have already developed first generation DMP and DSP tools and are actively upgrading these to comply with the FAIR principles.

Driving users will be all early implementers in other GO FAIR INs, such as [metabolomics, Seadatacloud, biodiversity, AGU, OPEDAS, OPERAS, IUPAC etc. etc.]. All scientists collaborating in these networks do not only engage in making legacy data FAIR but will also increasingly work on the FAIR publishing 'at the source' of de novo data in their fields. These include repositories (such as FigShare, Dataverse, Mendeley, DANS) and publishers (NPG, Elsevier, IOSpress, MITpress, Karger), but also the project officers of the associated early mover funders. These partners will rapidly and critically test tooling that has been developed in this

The basic schema for the tooling suite to be developed in this IN is explained here.

First of all in the picture below, the tool sets will directly support funders, researchers and repositories/publishers in the following process steps: **A**: funders who request FAIR compliant data stewardship plans will be provided with easy to use guidelines and tools that allow adaptation of requirements for metadata and other aspects at the 'individual Call/RFA level of granularity. **B**: tools based on (a.o.) DMP-online and the DS Wizard will assist PI's (and increasingly their professional data stewards) to construct a systemically structured and checked DS plan, that, at a minimum, meets the criteria set by the funder in de call for proposals. **C**: the plan can be submitted in one of the formats acceptable to the funder and where possible form a machine readable document itself. **G** The DStooling will also automatically suggest providers (such as repositories, FAIRfication services, journals/ puplication platforms and analytics) that are acceptable and/or certified by the funding organisation. The researcher will be presented with all acceptable options that lead to FAIR compliant data stewardship, so as to guarantee maximum freedom to implement and to prevent vendor-lock-in situations.


Findable Accessible Interoperable Reusable
H
execute with certified providers
G
I
can FAIR
reuse data and Services
receives funding, or Cloud Coins for DS
Application
APPROVED
E
PI
executes with certified steward
F
writes
or commits certified Data Steward
DS plan 'taken for granted'
A
D
submission to Funder
Research or Innovation proposal
C
approval by certified Data Steward
Data Stewardship plan
B
requests from PI

For all these services the FAIRwizard IN will rely on community adopted and GO FAIR best practice resources, thereby further driving convergence to broadly accepted open community standards, protocols, (meta)data models, ontologies, PID schemes and procedures.

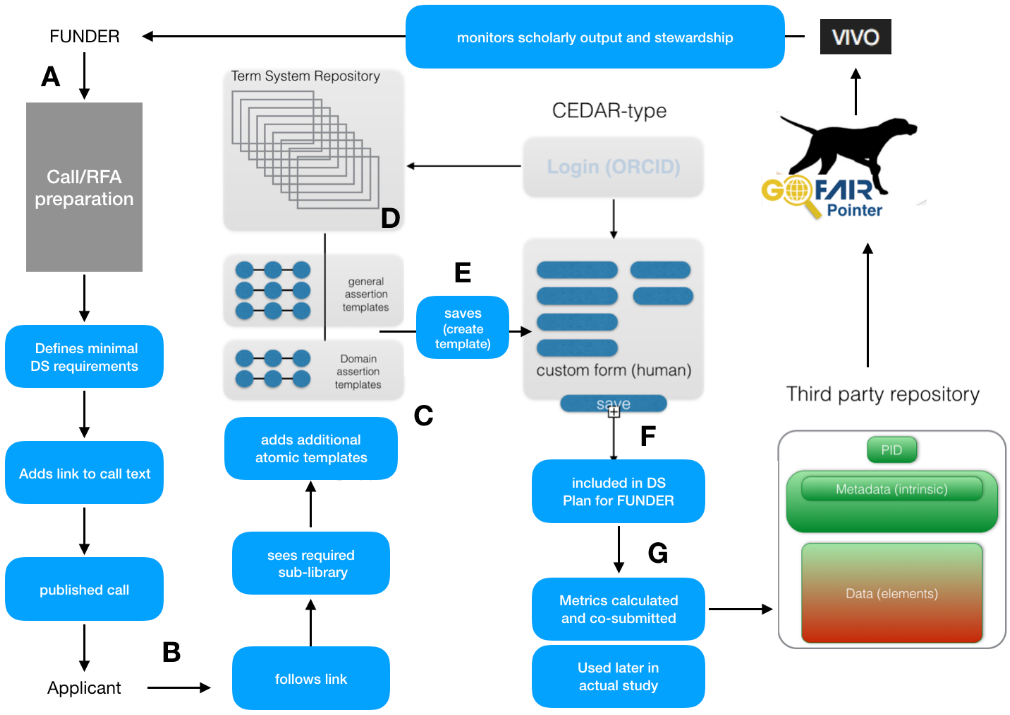
We will start with a very basic merger of mostly existing tooling in the institutions forming this IN. here we describe the functionality that this will enable in the form of a user scenario in the life sciences and health domain, but it should be emphasised again that the system will be expandable to any other domain, essentially by adding more terminology systems and ontologies/data models that are specifically covering that other domain.

User scenario:

The generic and domain-specific aspects of a good data (in fact research object) stewardship are intrinsically related to three elements: 1: the richness and quality of the metadata, 2: the perma-linking of these metadata to the actual research object, and 3: the perma-linking to scholarly information about the creators, the (re)users of the research object and their annotation and recognition behaviour. The latter also to allow optimal monitoring of the impact of the research objects on further research in order to enable proper award systems.
Therefore the scenario described here will largely focus on the requirements for the minimally requested metadata.

So we will combine our tooling and connect it to enable the following scenario:

A: The funding organisation (testers:HRB, ZonMW, NIH, NSF, ORFG) formulates a call for proposals (USA:RFA) and decides (potentially consulting professional data stewards where needed) on the **minimal requirements** for the research output to be stewarded according to the FAIR principles. As stated earlier, several high level decisions can be made without the proposed tooling to support the decision making process but these can still be made ‘custom’ requirements for each data stewardship plan that will be eligible under the call. These criteria can include for instance: Open Access publishing rules adopted by the funder, maximum allowable APC, list of accepted repositories (CoreTrustSeal for example), that support FAIR metadata and FAIR data publication, duration of data provision after completion of the project,


FUNDER
monitors scholarly output and stewardship
VIVO
A
Call/RFA preparation
Term System Repository
D
CEDAR-type
Login (ORCID)
GO FAIR Pointer
general assertion templates
Domain assertion templates
E
saves (create template)
custom form (human)
save
Defines minimal DS requirements
C
adds additional atomic templates
F
Third party repository
PID
Metadata (intrinsic)
Data (elements)
Adds link to call text
included in DS Plan for FUNDER
sees required sub-library
G
published call
Metrics calculated and co-submitted
B
Applicant
follows link
Used later in actual study

etc. etc. **B**: The funder provides a 'link' [button] in the online call text (please note, this happens outside legacy call management systems of the funding organisation, as it is an 'external link' in the HTML or PDF version of the call text. This link resolves to a (filtered) view on the nanopublication library described later, and leads the applicant to the metadata requirements (sub-library) that is **minimally needed to comply with the funder requirements**. **C**: The applicant, upon following the link, will first be presented with a human readable list of **human readable** 'nanopublication templates' [ref.]. These templates are defined at the 'semantic type level only' (for example [creator] [produced] [dataset]). And in fact these nanopublication templates are the '**atomic elements**' of a metadata template and the resulting file when the template is filled with concrete concepts (for example [ORCID] [produced] [DOI]. This 'atomic template' model is based on the premise that metadata are fundamentally treated as '**assertions about a research object**' (see also figure 1). These templates (if to be machine readable)[2] have the very simple format of 'subject-predicate-object triples', where the predicate is pre-filled and the subject and object can be filled by the user or pre-filled according to rules (for instance, the UPRI of the call that supports the selected projects could be pre-filled). Typically, generic nanopublication templates will have compositions like: [dataset]-[created by]-[ORCID1-n], [ORCID1]-[employed by]-[VIVO institution ID], [Geneset]-[created by]-[Methodology/instrument] etc. As a first step, the FAIRwizard IN will create a longlist of such templates with 'select/deselect' click boxes for the funder to preselect as 'minimal requirements to be captured and filled'. Obviously the researcher can select additional types of assertions to be added to the metadata, or preconceived options for structured annotation of the data (for instance of the type: [dataset]-[reused by]-[ORCID]. As all assertions will eventually be created as nanopublications they will ***themselves be adorned with the minimal essential provenance***, again as nanopublications, that trace the assertion back to the 'asserter', who is most likely initially the data creator or his data steward, but later on in the 'user defined' annotations of the dataset may be any researcher or annotator that wishes to 'assert' anything about the existing dataset (including for instance a critical appraisal of errors found in the data).

D: The preferred (and funder or GO FAIR approved) ontologies to be used (with existing autocomplete and mapping tools such as already available in for instance BioPortal and CEDAR) *will be enlisted* in applications such as FAIRsharing.org and will be **selectable by the user**. For instance, for the semantic type 'person', next to ORCID (preferred), older schemes such as SCOPUS ID (Elsevier), ResearcherID (TR, WOS), but also for instance a NARCIS-ID in The Netherlands, may be listed as acceptable ID's to use. For Genes, the NCBI and the EBI/ELIXIR identifiers will both be eligible, as these have been consistently and sustainable cross-mapped[3] and therefore machines and humans will not make mistakes about which gene is meant. For most of not all semantic types that a rich, machine readable metadata file will need to contain, proper ontologies already exist (FOAF, VIVO, Dublin Core etc….please add) and the institutes involved in this IN are either stewards of such core resources (for which sustainable funding is frequently lacking btw.) or they are in charge of creation, maintenance and expansion of these resources.

E: Once a final selection of all 'assertion types' in the list has been made, where a large percentage is likely to be 'standard and generic' throughout multiple disciplines and a subset will be specific for the type of research supported by the call (for instance 'metabolomics'), the system (*still to be developed but straightforward, based on the structured and nanopublication approach*) will **self-create, based on a chosen conceptual model,** a logically constructed, human readable and 'fillable' form as an aggregated **custom** template for the actual construction of the metadata.

[2] there can always be free text fields to enable rhetoric etc, but this relates to micro publication discussions [reference]. Nanopublications can be entirely machine readable, which is the subject of the memo.

[3] see for instance Bioportal and OLS

IN definition

F: These forms can be exported as part fo the DS plan and can be filled once the data are actually captured and processed. (very importantly, the IN will also use existing compost templates to 'extract' new atomic' templates from to be added to the library, so this approach entirely encompasses efforts already working with composite templates for domains). The unique feature of these forms (for instance produced by our CASTOR partner), based on the fact that they were originally created from a nanopublication library, is that, *once filled with the actual concepts in the 'blank fields' and saved, will automatically create both a human readable* ***and*** *a machine readable and actionable form of the intrinsic metadata, which is entirely FAIR compliant*. The same system can obviously also create 'annotation' templates in for instance CEDAR format that allow user defined secondary annotation of the research object. In addition, the system can be obviously used by anyone who starts creating a metadata template, including PI's at their own initiative, publishers, data and software repositories, or institutional data stewardship competence centers)

G: The forms will also (through integration of the DS wizard and the metrics) *generate an automatic 'score' on the level of FAIRness*. The forms can be both used as input for the DS plan to be submitted to the funder but also stored for actual use once the project is awarded and the data are generated for real (please not that this can all also be pertaining to code and other research outputs).

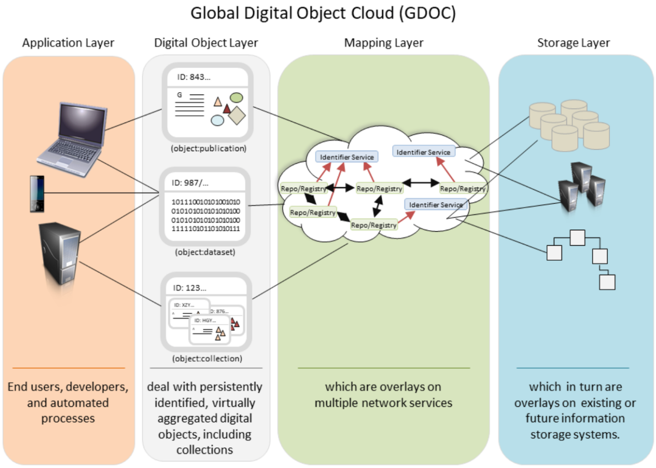
H: The templates will be automatically perma-linked to the 'ORCIDs' of the proposed data creators and stewards, and (via VIVO-type participants) to their institutions, which prepares for any tracking and recognitions activities later.
The publication of the metadata (and where possible the data elements themselves) in custom FAIRpoints will be offered. These can be submitted to any FAIR compliant repository or publisher. This will later allow funders and institutions to actually follow stewardship efforts, re-use of the digital assets coming from their funded research etc.

Risks:
One critical requirement for this distributed, Multi-stakeholder eco-system to work, and to be sustainable and scalable, is that multiple ontologies that have strong community uptake (minimally those used by other GO FAIR INs) are properly mapped, updated an sustained. This aspect is currently a weakness in the entire research infrastructure, which starts to be addressed by research infrastructures such as ELIXIR. These mapping tables are 'critical infrastructure' in the 'centre of the propeller'. Currently, however such services (such as or example BioPortal in the life sciences [more examples?] are built, maintained and funded largely by academic efforts and funded through volatile, few-year, cycles of public funding, frequently even in fierce competition with 'rocket science'. A key feature of GO FAIR as a movement is that we mobilise existing networks of excellence (gems) to converge and 'speak with one voice' to the funders (both public and private) of research and innovation about previously controversial issues such as the lack of sustainable investment in the 'rocket launcher' (the underlying infrastructure for 'rocket-open-science'). We recognise that the case for each individual service component (such as BioPortal, a single ontology, ISA tools, FAIRsgharing etc.) is difficult to make and is even more impeded by each of the academic groups running for the last possible funding source to keep the service up and running for another few months or years. It should be obvious that this is severe malpractice and may all by itself prevent the IFDS to develop rapidly unless we find a collective and sustainable solution. The first step we want to cover in this memo is to place these individual core resources in a much more comprehensive and internally consistent context. Mapping table, protocols and other community emerging standards should not only find a 'home' and a 'representation'/directory (such as for instance FAIRsharing), but should also be **collectively endorsed** and used in practice by much more coherent communities, which could be a key role of the GO FAIR INs [**'this is what we use'**]. A very important aspect of GO FAIR will be to support the process of coordination within and across implementation, training and certification networks to minimise reinvention of redundant infrastructure components, including such things as thesauri and domain-specific or generic ontologies, protocols, and other standards related elements of the IFDS. But, as said, we have learned that, traditionally,

domains operate in splendid isolation silos and that even within domains multiple standards, vocabularies, languages and approaches will continue to emerge. This is not only a nuisance and a lack of coordination and discipline, it is also an intrinsic part of the creative process that should be accepted and 'reluctantly supported' in order to further our knowledge and drive innovation. This means that 'mapping tables', 'libraries to choose from' etc. will continue to be crucial elements of the IFDS support infrastructure.

for consideration of interested readers only:

The Metadata layer needed in IFDS (C2CAMP IN).



This picture is taken from reference 2 (and the GO FAIR C2CAMP IN) and is a developing vision on a Global Digital Object Cloud and how such a Cloud (of which for instance EOSC should be part) will deal with storing, routing and (re-using) distributed digital objects.

The proposed FAIRwizard IN will help to consistently create metadata for all research outputs that enable the routing of data to tools and both to the needed compute to reuse the data and servies optimally for research and innovation.

Proposed partners in this IN

- **Health research Board** (Irl), early mover in FAIR DS
- **ZonMW** (Dutch health research funder (NL), early mover in FAIR DS
- **LSH-Health Holland** (early supporter of FAIR for health research)
- NIH (BD2K, Commons, Interagency coordination group USA)
- NSF (Interagency coordination group)
- **Digital Curation Centre** (UK)- DMP online and vast experience in data curation aspects (and partner in USA)
- **Dutch techcentre for the Life Sciences** (Mind Map for data stewardship, ELIXIR node and supporting ZONMW
- **ELIXIR Node Czech Republic** (DSWizard, metrics)
- **FAIRsharing**
- **CEDAR/BioPortal**
- **National Data Service**
- **DANS** (RDA node)
- **DATAverse**
- **VIVO** (Scholarly record)
- **CONVERIS** (using VIVO ontology)
- **PURE** (using VIVO ontology)
- <> **OPEDAS IN**
- <> **METRICS IN**
- <> **AGU IN**
- <> **OPERAS IN**
- **<> SEA DATA CLOUD IN**