Data Needed to Identify Plan S Compliance

Commissioned by Jisc on behalf of Science Europe

This report was commissioned by Jisc on behalf of Science Europe.
Please direct queries to info@coalition-s.org.

First draft submitted: 20 December 2019
Revision submitted: 14 January 2020
Final Version submitted: 24 January 2020
Final Version with feedback incorporated: 6 February 2020

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report was commissioned by Jisc on behalf of Science Europe. It is the result of a project which examined:

- the data needed for authors to identify Plan S-compliant publication venues;
- open questions about data needed for compliance as of December 2019;
- the readiness of key data sources to provide the data; and
- a view of gaps in the short and medium term, and how to fill them.

The project's scope covers the **data** needed by a third party to produce an author-facing tool that will allow an author to identify publication venues before submission. The tool itself is out of scope, and is being looked into via a separate project running parallel to this one. This report and its accompanying analysis are intended to feed into the tool's Statement of Requirements (or similar). The Appendix – Project Terms of Reference recaps this report's terms of reference. A Project Steering Group comprising representatives from Jisc and cOAlition S oversaw the project, and included Science Europe as an observer.

Plan S requirements indicate a clear aspiration, but not all are sufficiently detailed in providing for a technical specification. In analysing the data requirements, we therefore encountered questions about some of the details behind them. We used our regular contact with the steering group throughout the project to clarify what approaches we should use, or to capture open questions. We had a limited number of hours available for the work, so we agreed a prioritised list of stakeholders we interviewed as part of our investigations. This report represents the results of our discussions, the process of analysis we undertook, and response to the steering group's request for our independent views. The work was undertaken in the last quarter of 2019, and so its results represent a snapshot of activities at that time. Details about Transformative Journals, and Information Power's report about price transparency were published during the course of the project. The implementation of Plan S continues to evolve.

The report consists of five major sections. "Assumptions and Questions" summarises the results of our discussions with the Project Steering Group, covering recommendations and questions arising. The next two sections are technical: "Analysis of Data Needed" identifies common threads and generic structures; the "Data Specification" section goes into specific details of the data. The specification should be read in conjunction with the detailed spreadsheet accompanying this report, *JISC Plan S Data Spec.xlsx*. "Assessment of Sources" examines what data sources are available, based on the priorities we were given. "Conclusions and Recommendations" presents our requested views on next steps.

A summary of our findings is as follows.

The data about compliance should allow multiple levels of detail for a given publication venue. As well as indicating whether a venue is Plan S compliant overall, the data structures capture how it measures up for each of the four routes to Plan S compliance, and in turn how each specific requirement contributes to each route. For example, a journal may be compliant via the fully OA route because it is in the DOAJ, has appropriate editorial policies, offers the correct licenses, and so on. A tool built on this data structure then has flexibility in how much detail it presents to the end user.

No data sources currently include **all** the data needed to determine plan S compliance. Some key requirements can therefore not be measured without further work, and some are ambiguous. We suggest that cOAlition S should take a phased approach to enforcing requirements where data sources are currently unworkable:

1. No industry-standards or sources exist for information about publishing statistics.
2. No industry-standards or sources exist for information about publishing prices and costs. The Plan S requirements are unclear about exactly what information is required. (Although we note the work by Information Power and the Fair Open Access Alliance.)
3. Requirements specifying "in the process of being registered" in the DOAJ or OpenDOAR would not be workable in practice. The sources do not implement such a process, and handling rejections may prove complicated. Our discussions suggested that these requirements were anticipating a surge in demand, so might better be addressed by ensuring the data sources have sufficient interim resources to manage demand.
4. Requirements stating "at no additional cost" do not specify the baseline against which the cost is calculated.
5. Formats for metadata are not specified for several requirements where things like PIDs, "quality metadata" and "machine readable metadata" are mentioned. The data specification therefore simply flags absence or presence of metadata. However, without standards in place such metadata may offer limited value. We suggest that priority should be given to specifying a limited taxonomy for license information embedded in articles.

Given the tight implementation timescales desired by cOAlition S, we ratify the Compliance Task Force's approach of nominating a few key data sources with a view to scaling them.

- The approach of multiple whitelists is an efficient way to analyse the key publication venues. If a publication is present in a whitelist, and passes various checks, it can be deemed compliant. Its absence implies non-compliance, without need for further data.
- Curation of the data is delegated to the whitelist operator, with cOAlition S trusting the operator's judgement.
- We recommend that cOAlition S quickly clarifies its policies and priorities with whitelist operators, and works with them to make resources available to cover any gaps.
- In order to balance rigour against prohibitive costs, we recommend a mix of proactive publisher deposition of compliance data into the whitelist(s), complemented by random spot checking by the whitelist operator to verify accuracy.
- We recommend running a focus group involving whitelist operators and publishers to clarify the best balance between voluntary or mandated deposition of compliance data, responsiveness and rigour of data validation. The results could be used to set expectations and foster understanding between all stakeholders.
- We recommend that cOAlition S produces a draft timeline for phasing in requirements that are not prioritised for the tool's initial launch, so all stakeholders have clear expectations and can make appropriate plans.
- We have focused on the data specification here. However, data ownership and governance must be considered. For each route we suggest that it is important that one source only is deemed to have authority and offer a "single version of the truth" allowing for unambiguous compliance assessment. In principle, the requirements for

open licenses mean that any data collected could be transferred to alternative providers in the future.

- We recommend that cOAlition S decides on the following before inviting tenders for the compliance checking tool: which mandatory requirements are needed for launch; policy details about mandating compliance data deposition (or not) and data verification; rules for multi-author papers and handling policy exceptions.
- We anticipate the following details would be handled by the tool's developer: the process for escalating and resolving questions about the data; details of engagement with data providers, end users and publishers (if applicable); data update frequency and processes; specific metadata taxonomies.

The key data sources (whitelists) can be analysed by Plan S compliance route. cOAlition S has made clear its need for speed of implementation, so we have prioritised the most mature data sources in our assessment. Timing is already tight for 2020 implementation, so we also recommend that cOAlition S quickly agrees budgets and expectations with the key curated sources (e.g. DOAJ, Sherpa, ESAC), so they can proceed with any necessary implementation.

- The DOAJ is the clear choice as a whitelist for fully OA journals. It is mature and robust, and the team are already working on plans to add details for Plan S. Further analysis is needed to estimate an anticipated spike in registrations, and agree how this is best addressed.
- Sherpa (RoMEO and OpenDOAR) is the clear choice for whitelists for the Subscription/Repository route. (Other sources exist, but have significantly less coverage.) cOAlition S would need work with Jisc to agree priorities, address issues of perceived unresponsiveness, and make relevant data available under CC0 licences. (Note that very few of the data requirements for repositories are currently tracked by anyone. Data about Repositories was de-prioritised during the course of this project.)
- A centralised database of Transformative Agreements (TAs) needs to be built, to map agreements to institutions and individual journals. Note the difference between curation and collation. We discussed that individual consortia should **curate** their own agreements with their suppliers, and be responsible for ensure up to date accurate lists of applicable journals. (So, in essence, each consortium maintains its own whitelist.) A central database would then **collate** the locally-curated data into a central resource. ESAC currently tracks only data for the agreements as a whole. A "Plan S compliant" indicator is not currently implemented. A database to resolve to the individual journal level for each TA and institution would require significant extra work. cOAlition S would need to work with a provider (e.g. ESAC, or the Netherland's SURFmarket) to specify the work needed, and agree resourcing.
- Likewise, a centralised database of Transformative Journals (TJs) needs to be built. We discussed that cOAlition S might curate an approved list. Adding an indicator per-journal to RoMEO might be a logical starting point from which to collate the results. cOAlition S would need to work with Jisc (or other 3rd parties) to specify the work needed, and agree resourcing.

ASSUMPTIONS AND QUESTIONS

All stakeholders we consulted agreed that some items in the Plan S requirements are directional, but lack enough detail to form a detailed specification. Indeed, the point of this investigation is to identify gaps and map out how we might fill them in a pragmatic way. For example, many requirements specifying metadata or machine readability do not specify the exact standards or taxonomies these would use.

Below, we note the approaches taken which guided our data specification, followed by open questions that need to be resolved by the larger Plan S Compliance Taskforce.

MAIN REQUIREMENTS

During the course of the project, our investigation and analysis highlighted questions about priorities and definitions.. In general, the philosophy behind our recommendations is to avoid the perfect becoming the enemy of the good.

GENERAL LEVEL OF DETAIL

Over the course of the project, we confirmed the following general approach about the level of detail that the data should address.

1. A tool using the data will need to deliver both:
    a. an **Overall Indicator** of "compliant/not compliant" (or "yes/no") which applies to a publication venue, overall and **for each Compliance Route** separately; and,
    b. an indication of the **Contributing Requirements**, as a series of "yes/no" indications indicating whether each of the individual Plan S requirements are met (e.g. the journal has the right license, the right editorial standards, etc.)

    Each contributing requirement has to be met for a publication venue to be deemed to be "compliant" for a given route. As long as the venue is compliant for at least one route, then it can be deemed to be Plan S-compliant overall.
2. We assume descriptive information about specific requirements is **optional**. For example, consider the requirement "The journal/platform must provide, on its website, a detailed description of its editorial policies and decision-making processes." To be compliant, the journal either provides the information or it doesn't ("yes/no"). However, one would assume that capturing the URL of the information would be needed by those certifying the journal. So, in this case we would specify a related URL ("descriptive information") as optional data.
3. If a Plan S requirement's details are ambiguous, we specify a placeholder data point. E.g. "Does journal x have information about costs on their website? (yes/no)" – is the best indicator we can provide absent (at present) a breakdown of specific costs.
4. We focus on the underlying data that could be incorporated into a tool, solely to identify the journal's compliance. We do not include other journal metadata that might be considered useful (such as subject area).

PRIORITISATION OF REQUIREMENTS

Discussion with the Project Steering Group suggested that we should focus only on **Mandatory** Plan S requirements for this study.

We will further split the mandatory requirements into :

1. The "**most important**" Plan S requirements – i.e. those for which reliable data are available.
2. "**other**" requirements – i.e. those where exact details or standards have yet to be defined, such as information on cost breakdowns.

Prioritising the "most important" information is a means to square the reality of what is realistically available now, with the aspirations for a comprehensive suite of information in the future.

- Where reliable data sources do not currently exist – for example, the publication of detailed breakdowns of costs – journals will **default to be deemed compliant** in these areas.
- Over time, as data sources improve, some / all of the indicators from "other requirements" will move into "most important" and journals will need to update their offerings to remain in compliance when this happens.
- Note that this approach prioritises Mandatory Plan S requirements. It is not related to Recommended requirements in the Plan S guidelines.
- The Compliance Task Force will need confirm its position on specific requirements; the results of this study can help guide its decision. The list for discussion is shown in the "Open Questions" section below.

The Compliance Task Force further agreed that Repositories will be out of scope for the author compliance tool, as authors may not be able to determine details about these at the time of submission.

AUTHORITY OF TRUSTED SOURCES

During the course of the project, we raised some questions about how to handle aspects of data policy with the DOAJ. The conversations were timely and relevant, as the DOAJ are a nominated source and were in the process of analysing details for providing Plan S compliance information. The Project Steering Group provided the following answers**, which could apply in principle to any nominated 3rd party providing data curation.**

1. How far does checking [of the DOAJ] have to go? What proof (if any) is needed to make sure journals are doing what they say they are doing?
    a. The DOAJ's current check seems adequate.
    b. cOAlition S would trust trustworthy services to do all that is reasonable.
2. Will cOAlition S need external access to underlying [DOAJ] data, or is the data they publish sufficient?
    a. No; access is not needed. If (say) a funder queries a journal's inclusion, we anticipate that this would be resolved via a discussion on a case by case basis.

    b. (**Note**, this therefore implies that any 3rd party providing data must have some sort of process to gather and resolve queries about its data.)
3. To what degree do cOAlition S trust the DOAJ's judgement where criteria are ambiguous, or where assessment is a matter of judgement? Should DOAJ wait for precise details of all criteria or can they make recommendations?
    a. The DOAJ team will separately raise questions inviting cOAlition S to express views on specific criteria that, without clear definition, it is not possible to resolve (or about which the DOAJ would need to make its own judgement. Delta Think is doing the same as part of this analysis.
4. Are there any unacceptable answers to requirements…e.g. unacceptable performance levels?
    a. Plan S's focus is on transparency, not performance assessments.
5. Can we consider a staged release to phase in all requirements over time?
    a. It was acknowledged that this is a possibility.
    b. Hence the "Most important" vs. "Other" requirements approach in the outline above.
6. How do we handle "in the process of being registered"?
    a. …given that no mechanisms exist? Further, how would we handle cases where a journal or repository applies but is subsequently rejected…would compliance then be withdrawn? And what would happen to authors who had submitted in good faith in the interim?
    b. We discussed that the aim behind this was to handle an anticipated spike in registrations, and the lead times of working through any resulting backlog.
    c. The notion of "in process" is therefore something that will not be supported.
7. What about strongly recommended requirements?
    a. Not needed yet – focus on mandatory ones.

OTHER ASSUMPTIONS

These are some detailed assumptions we made about mandatory requirements.

Requirement		Notes
III-1.1.1	**Basic mandatory conditions for all publication venues…** Pre-requisite: we must be able to identify the journal. Ideally, we would specify ISSN-L, but they do not offer complete coverage. We assume any validated ISSN as a bare minimum.	
III-1.1.2	cOAlition S emphasises the need for high quality journals, therefore requiring journals/platforms to have a solid system in place for review according to the standards within the relevant discipline and guided by the core practices and policies outlined by the Committee on Publication Ethics (COPE). Details must be openly available on the respective journal and platform websites. In particular, payment of publication fees or waiver status must not in any way influence the editorial decision-making process on the acceptance of a paper.	Claimed COPE membership does not imply practices are followed. Note discussion under Authority of Trusted Sources above - we assume trusted data sources confirm quality. The data therefore simply indicate compliance ("yes/no"). We assume that as items must be peer-reviewed, simply depositing in a Repository would not be compliant.
III-1.1.7	Use of persistent identifiers (PIDs) for scholarly publications (with versioning, for example, in case of revisions), such as DOI (preferable), URN, or Handle.	We only flag presence of adequate metadata here. No PID taxonomy is defined.
III-1.1.8	Deposition of content with a long-term digital preservation or archiving programme (such as CLOCKSS, Portico, or equivalent).	Assume we do not need to specify the repository.
III-1.2.5	The journal/platform must provide APC waivers for authors from low-income economies and discounts for authors from lower middle-income economies, as well as waivers and discounts for other authors with demonstrable needs. Waiver policies must be **described clearly on the journal website/platform** and statistics on waivers requested and granted must be provided.	URL assumed to be compulsory. Assume we don't need info on programme memberships (e.g. Hinari or R4L). No definitions of statistics format is specified.

OPEN QUESTIONS

Here we call out the key decisions that need to be taken by cOAlition S or its representatives to round out a full data specification. Our approach of using generic "yes/no" indicators of the various requirements means that we do not anticipate the data needs being dependent on the answers. Deciding what constitutes a "yes" or a "no" is a matter of policy, not data design.

GENERAL GUIDING DECISIONS ON DETERMINING COMPLIANCE

a. **What are the most important pieces of information needed to determine compliance? How might we phase in those where data sources are not mature enough, or requirements sufficiently well-defined?**
    i. As discussed above, we need to be mindful of the practicalities of which information is available and defined.
    ii. To facilitate this decision, the next section below, "Requirements Needing Prioritisation," outlines our analysis of which requirements do and don't have immediately available, reliable data.

b. **How do we handle unspecified metadata details in requirements?**
    i. E.g. "Full text [must be] stored in a machine-readable community standard format **such as** JATS XML."
    ii. **By default, the data specification simply indicates the presence or absence of metadata in these cases.** At some point, cOAlition S will need to decide: What would constitute acceptable formats or standards? Is the intent that they need to be interoperable with each other – in which case, which standard taxonomies, etc., would be needed for each?
    iii. If the intent is to address future-proofing against gathering statistics for monitoring, and making tools more usable, **we recommend that cOAlition S facilitates coordination of standards between those implementing tools**.
    iv. One notable priority is listing licenses, as acceptable ones are clearly defined. **We recommend that a taxonomy and structure is specified for including license information in articles.** Given the predominance of CC and other licenses, this should prove to be a workable solution. We recommend sense-checking feasibility with some publishers prior to specifying a policy.

c. **How do we resolve conflicting policies for multi-funder papers?**
    i. This is assumed to be a decision for policy makers on a case-by-case basis.
    ii. Whoever implements specific tools would need to determine the best User Interface to handle this; it may require authors to liaise with each other and lie outside an abstract specification.
    iii. The feedback to this document's initial draft included a discussion about this. However, it is an issue of policy and agreements, separate to a data specification.

d. **How do we handle funder exceptions? (E.g. allowing use of CC BY-ND?)**
    i. It is important to keep focus: the tool focuses on general policies **per journal**. Funder exceptions are made on a case-by-case basis per article.
    ii. Our discussions suggested that it would therefore not be realistic for a tool to fully compute outcomes for a specific paper. Results would need to be presented as

general rules for a journal, and authors would need a clear method to raise and resolve questions specific to their funder and paper.

e. **Should the compliance routes "cascade"? Does each journal need to be analysed across each of the four routes of Plan S compliance?**
    i. Feedback to the draft report questioned our approach to analysing each compliance route separately. (See "Routes to Compliance" below.) For example, if it's a fully OA journal then nothing else is needed.
    ii. **Our response:** from a data perspective, we need to specify all routes. The Plan S requirements do not state a hierarchy, so we cannot establish a set of rules for cascading.
f. **Should publishers be formally mandated to keep data sources up to date? Do requirements specifying the provision of metadata imply any level of quality?**
    i. Do requirements for DOAJ and OpenDOAR registration imply any responsibility on the part of the publishers to proactively deposit information, or should the registries in question be solely responsible for gather information?
    ii. The requirement to specify metadata is meaningless if the data is of poor quality or not interoperable.

TRANSFORMATIVE AGREEMENTS

TAs present particular challenges as they allow for temporary exceptions, and involve multiple stakeholders and data sets. During discussion with the Project Steering Group the following definitions and methods were discussed. These need confirming through a cOAlition S policy decision.

a. **For a given TA.** In general, each cOAlition S member should decide if its agreements are compliant. Notes:
    i. Each cOAlition S member should decide if the agreements relevant to it are compliant and provide a simple "yes/no" indicator.
    ii. **How do we handle separate funders?** Agreements are enacted (typically) between a consortium and a publisher. Predominantly national funders (usually) have an obvious consortium negotiating relevant TAs (UKRI and Wellcome -> Jisc. NWO -> UKB). But Funders are not always involved – e.g. Gates or the EC. So: **Can we assume that funders will offer "yes/no" indications of TA compliance? …if so, who will maintain and curate such a list?…if not, how do we determine which information applies and which journals are covered where there is no direct agreement in play?**
b. **For a specific Journal** within an agreement. A journal is deemed "TA compliant" if:
    i. An Agreement is verified by Consortium or Funder as being compliant (as above)
    ii. Authors can participate in it (as defined by their affiliated institution)
    iii. The Journal is part of the Agreement.
    iv. The Agreement is active when the journal is checked.
c. **Data ownership.** From a database perspective, confirm that the **consortium** should hold the definitive list of participating institutions and applicable journals. Notes:
    i. Who makes the list available depends on the relative abilities of the consortium's systems.
    ii. Publishers may hold this information, but in practice it may not be accurate. It may also not be appropriate: should the buyer be deemed to be accountable in

principle, even if it chooses to work with publisher partners? (Although note, in some cases it is the opposite: publishers offer both accurate information and useful services to help a consortium determine compliant publishing options.)

iii. Ideally, local (per-consortia) data about institutions and journals would be collated in a central database tool. **No such resource currently exists, so how should this be addressed?** (Note that ESAC does not currently cover this level of detail.)

TRANSFORMATIVE JOURNALS

The Project Steering Group flagged the need to address these after the project started, but the open consultation is on-going and due to close on 6 January (just after the draft report was submitted).

a. **Confirm: Who will decide if a given journal is compliant?** The relevant Coalition S committee (to be decided)
b. **We will assume The Plan S Secretariat will address**:
    i. A need to maintain machine readable list of journals which are deemed to be transformative
    ii. "Compliance" to be defined by a journal's presence or absence in the list
    iii. Add placeholder for the data
    iv. We assume a single list, and single authority
c. **Confirm: a journal is deemed compliant simply if it is a member of the list at the time it is checked.**

REQUIREMENTS NEEDING PRIORITISATION

This table summarises requirements for publication venues that are ambiguous, or for which reasonably comprehensive data is not available.

Requirement		**Notes, Assumptions and Questions**
III-1.1.3	The journal/platform must provide, on its website, a detailed description of its editorial policies and decision-making processes. **In addition, at least basic statistics must be published annually,** covering in particular the number of submissions, the number of reviews requested, the number of reviews received, the approval rate, and the average time between submission and publication.	We assume data source will assess this ("yes/no"), and a URL is needed pointing to the description. **No standards or mature sources exist for statistics**. Although we note the approach proposed by Fair Open Access Alliance (FOAA) in response to the call for tender to build the "Open Research Europe Publication Platform." Publishers may be unable to provide statistics about reviews due to confidentiality concerns.
III-1.1.4	The journal/platform must accept the retention of copyright by the authors or their institutions, **at no extra cost.** …	**Unclear how to determine "at no extra cost"** - and extra cost compared to what baseline? E.g. is charging APCs for different licenses not allowed?
III-1.1.5	The journal/platform must either enable authors to publish …, or to deposit the AAM or VoR in an Open Access repository **at no extra cost** …	**Same question about "extra costs" as for previous requirement.**
III-1.1.9	**High-quality article level metadata** in standard interoperable non-proprietary format, under a CC0 public domain dedication. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).	The data spec only flags presence of adequate metadata here. No standard or taxonomy is defined. We assume indicators for each required field need not be added (e.g. funder name, grant number/ID). Note: other metadata covered below.

III-1.1.10	**Machine-readable information on the Open Access status and the license embedded in the article, in standard non-proprietary format.**	The data spec only flags presence of metadata. There is no industry-standard form for embedding status (is this open: yes or no?) or license information into articles. NISO Access License and Indicator is NOT sufficient (it only indicates "free to read" + a license URL - no machine readability of license is guaranteed at the URL). Consider Unpaywall lists? Should a taxonomy allow non-compliant options for auditing reasons for non-compliance?
III-1.2.1	The journal/platform must be registered in the Directory of Open Access Journals (DOAJ) **or in the process of being registered**.	We assume a DOAJ identifier is not needed (or it could be a URL to an entry)? **"In the process of"** - the DOAJ does not support this – refer to item 6 under "Authority of Trusted Sources" above. We assume "in process" not to be needed.
III-1.2.4	Transparent costing and pricing: **information on the publishing costs and on any other factors impacting the publication fees** must be openly available on the journal website/publishing platform (see also Part II Section 5 above).	We only flag presence of metadata. The list of exact data needs specification. [What are the factors or services that should be covered? Does this cover prices (APCs charged by publishers), or costs to the publisher, or both? No data source currently exists to capture this. No industry standards exist, although we note the FOAA approach as mentioned above (III-1.1.3 in this table), and the analysis undertaken by Information Power (published just as this report was submitted).
III-2.1.1	The repository must be registered in the Directory of Open Access Repositories (OpenDOAR) **or in the process of being registered**.	(Same issue for OpenDOAR as for the DOAJ; see III-1.2.1 above.)

AMBIGUOUS METADATA

The following table lists the mandatory requirements for publication venues where the specified metadata are subject to interpretation…

III-1.1.1	Basic mandatory conditions for all publication venues **(as indicated previously, the need for a venue identifier is implicit)**
III-1.1.3	The journal/platform must provide, on its website, a detailed description of its editorial policies and decision-making processes. In addition, at least basic statistics must be published annually, covering in particular the number of submissions, the number of reviews requested, the number of reviews received, the approval rate, and the average time between submission and publication.
III-1.1.7	Use of persistent identifiers (PIDs) for scholarly publications (with versioning, for example, in case of revisions), such as DOI (preferable), URN, or Handle.
III-1.1.8	Deposition of content with a long-term digital preservation or archiving programme (such as CLOCKSS, Portico, or equivalent).
III-1.1.9	High-quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).
III-1.1.10	Machine-readable information on the Open Access status and the license embedded in the article, in standard non-proprietary format.
III-1.2.4	Transparent costing and pricing: information on the publishing costs and on any other factors impacting the publication fees must be openly available on the journal website/publishing platform (see also Part II Section 5 above).

Repositories have similar issues for the requirements for Use of PIDs, High quality article level metadata, and machine readable information on the Open Access status and article license.

ANALYSIS OF DATA NEEDED

Indicating Plan S compliance requires assessing multiple data points. This means we need to identify the data needed and design a "recipe" (aka an algorithm, or series of logical steps) to combine the multiple points. We need first to analyse how the different compliance routes and criteria fit together in order to determine the appropriate data points.

Our focus here lies with the underlying data and indicators. It will be up to those implementing the tools used in practice to build out User Interfaces on top of the data, e.g. to present authors with a list of journals to choose from (a pick-list), or a search facility, and present the results in a suitable format for each.

In this section we analyse the Plan S rules and the User Stories we were asked to assess, to understand how things fit together, and identify common threads and generic structures. The "Data Specification" section below then goes into specific details.

ROUTES TO COMPLIANCE

At the time of this investigation, there were FOUR stated routes to Plan S compliance, with combinations of multiple data points needed to indicate compliance for each. For clarity, we refer to the Plan S routes (as summarised in Figure 1, below) using abbreviated terms as follows:

1. **Fully OA**, covering "Open Access Publishing Venues (journals or platforms)" which we take to include Fully OA journals. Note: this includes all Fully OA journals, regardless of the payment mechanism – so it covers both "Gold" and "Diamond".
2. **Subscription/Repository**, covering "Subscription venues (Subscription/Repository route)" journals that allow deposit in Repositories and the Repositories themselves.
3. Transformative routes, covering "Transition of subscription venues (transformative arrangements)," which we further divide into:
    a. **Transformative Agreements (TAs)** – including any journals covered by a Transformative Arrangement (TA). The expectation is that these will be hybrid, but an agreement could technically include any journal.
    b. **Transformative Journals (TJs)** – covering journals as defined by the update published during the course of this investigation.

Figure 1 – Original Plan S Compliance Routes – note the Transformative Journals have since been added.

All scholarly articles that result from research funded by members of cOAlition S must be openly available immediately upon publication without any embargo period.

There are three routes for being compliant with Plan S:

	Open Access publishing venues (journals or platforms)	Subscription venues (repository route)	Transition of subscription venues (transformative arrangements)
Route	Authors publish in an Open Access journal or on an Open Access platform.	Authors publish in a subscription journal and make either the final published version (Version of Record (VoR)) or the Author's Accepted Manuscript (AAM) openly available in a repository.	Authors publish Open Access in a subscription journal under a transformative arrangement.
Funding	cOAlition S funders will financially support publication fees.	cOAlition S funders will not financially support 'hybrid' Open Access publication fees in subscription venues.	cOAlition S funders can contribute financially to Open Access publishing under transformative arrangements.

For any chosen route to compliance, the publication must be openly available immediately with a Creative Commons Attribution license (CC BY) unless an exception has been agreed by the funder.

Our analysis suggests that each compliance route should be handled separately, with compliance options then being a sum of those available for all four routes. E.g. a Fully OA journal (by definition) allows deposit of an article, so is technically both Fully OA compliant and Subscription/Repository compliant. Or a Hybrid journal may allow deposit (Subscription/ Repository compliance) AND fall within a TA covering the author (TA compliance) AND be recognised as a transformative journal (TJ compliance).

So, we will conceive of a "**compliance pathway**" for each compliance route, which combines the data, an algorithm (i.e. logical rules) to determine compliance, and identifies key data sources. The data specification must therefore include a relevant series of metadata indicators for each source and specify the decision logic to deduce compliance.

A tool built on this data model could then demonstrate an appropriate level of detail to its users, from a simple "yes/no" to a breakdown of reasons why a venue is compliant or not.

THE "AUTHOR" USER STORY

The terms of reference suggested that we prioritised the "author" user story. Our analysis and feedback from experts suggested this was a logical starting point, and could serve as a building block for the other user stories. So, we start by looking at the data needed for an author to determine their Plan S compliant publishing options.

Handling multi-author papers lies out of scope for this analysis. Our data model allows an author to determine the Plan S compliant venue(s) for them.

We assume any tool that is built will operate as follows. The user [author] enters a funder, institution, and journal ISSN for which they wish to determine compliance, and should reasonably be expected to know this information. In response, the user is shown a Compliance "yes/no" answer for each compliance route, and, optionally, the contributing requirements.

THE SINGLE JOURNAL

Technically, Plan S requirements are the property of an article. However, our scope here is to look for publication options, and articles are published in Journals (or journal-like entities). Plan S requirements for minimum editorial and license requirements imply a journal's involvement, and specifically refer to journals (or journal platforms).

Without a journal's involvement, in practice there can be no editorial policy implemented or license granted. Any outliers that are not technically journals may theoretically exist, but they would exhibit journal-like functions and so might reasonably be considered to be journals. Journal requirements also apply to all compliance routes. Fully OA is self-evident (by definition); TAs and TJs are predicated on named journal lists. The Subscription/Repository route requires an edited paper to be deposited, even if the choice of repository is subject to separate criteria.

So, our data specification uses the **Journal** as a building block. Our suggested approach is to conceive of the Author User Story as identifying if ONE journal is compliant. If we can determine compliance for one journal, then we can infer compliance for an arbitrary list of journals.

The data must allow each journal to be analysed across each of the four routes of Plan S compliance.

FULLY OA

If a specified journal is fully OA, and it meets Plan S criteria, then we can determine it to be Plan S compliant via the Fully OA route. This is the simplest case, and can be inferred from the journal's ISSN.

SUBSCRIPTION/REPOSITORIES

If a journal allows deposit of an AAM or VoR in a repository, then the **Journal** is compliant via the Subscription/Repository route. But, an author would then need additional information to determine which **repositories** are compliant to fully determine their compliance options. Note, Fully OA compliant journals are Subscription/Repository route by virtue of their licenses.

As with journals, if we can determine compliance for one repository, we can scale this to a list of them. This might be a list such as OpenDOAR, which the author needs to check separately. Or, where a journal for funder policy specifies named repositories, the data would need to capture the named list. A tool could use this list to automatically check repositories specific to a given journal.

TRANSFORMATIVE AGREEMENTS

Identifying compliance via TA (for the Author use case) requires a combination of data points:

1. A means of identifying an applicable Agreement.
    a. TAs exist between publishers and institutions, for example in "Publish and Read" deals.
    b. Identifying a TA will require information about the journal in question (from which a publisher can be inferred), the institution in question and (optionally) the funder of the research.
2. A date range across which each TA applies. For the sake of simplicity, we do not examine the specifics of submission vs. publication dates. A tool might assume the current date, or allow an author to specify an intended submission date (or other date of interest).
3. A list of journals across which each TA applies.

We assume that if a journal is present in a compliant TA, then that journal is compliant. The point of TAs is to offer temporary exemptions for journals that would otherwise not meet other compliance routes. So, given identifiers for a journal, institution, and funder; a list of TAs; and a list of journals present in each TA, we can assess an individual journal's TA compliance.

We recognise that this may lead to more onerous requirements being placed on cOAlition S's preferred fully open journals. However, our understanding was that this approach would a valid interpretation of Plan S Requirements. Further, to impose additional requirements on TAs risks making them unworkable in a realistic timescale, as the needed data sources are simply not available.

TRANSFORMATIVE JOURNALS

Transformative Journals are a new concept, and their requirements were added late in this project. For the purposes of a data specification, we assume that some sort of approved list of applicable journals will be maintained. If a journal is present in the approved list, then it can be deemed to be compliant.

CORE IDENTIFIERS

The building blocks above presuppose that we can unambiguously identify journals, repositories, institutions, funders and TAs. In theory, any reliable identifier could be used for these, although in practice mixing different standards will pose implementation problems. So, we suggest the following **core identifiers**, which an author needs to specify to determine

the compliance of a given journal. We suggest sources as per the project's scope, to build on work done so far, and to leverage emerging *ipso-facto* standards.

Entity	Identifier	Source	Notes
Journal ID	Validated ISSN	issn.org	Uniquely identifying journals is a challenge as we lack a universal, reliable identifier. The ISSN exists, but journals may have multiple ISSNs (e.g. print, electronic). In theory, the ISSN-L (Linking ISSN) should unambiguously resolve these. But in practice, its data quality and uptake varies. Curated sources will need to validate journal ISSNs and tools will need to accept any ISSN. ISSN-L tables are available for free download from issn.org to help with disambiguation, but their license is not clear.
Repository ID	OpenDOAR ID	Jisc	OpenDOAR is the most comprehensive database of repositories. (But note issues with its non-CC0 license as noted below.)
Institution ID	Email address domain name	-	There is no universally-used identifier for Institutions. The context here is TAs, and locally-maintained lists of participating institutions. Our research suggested that the domain name of the author's email address would be the most accessible identifier. If a curated taxonomy was needed, the GRID database from Digital Science is comprehensive and open.
Funder ID	FundRef	CrossRef	There is no universally-used identifier for Funders, although the FundRef identifier is an emerging standard. Again, the context here is TAs so it should be possible to leverage the FundRef ID for the relative few TAs that will apply.
TA ID	ESAC ID	ESAC	This assumes ESAC will add a "Plan S" compliant flag to their data. Data available under CC0.
TJ ID	Validated ISSN	TBD	Some list of approved Transformative Journals would be needed. No such list exists.

SCALING UP

The exact specifications of the operation and UI of tools lie out of project scope. The key principle is that a tool determines a list of core identifiers as specified here, and can use the logic outlined in the sections below to identify venues in various curated sources of information and determine compliance. In principle we can then scale across our use cases as follows.

1. An author-facing tool could allow an author to choose from an arbitrary list of journals. (A tool might to be pre-loaded with lists of (say) journal names and their identifiers to facilitate user-friendly choices. The details would lie with the tool's developers.)

2. It would then infer a list of journal identifiers (ISSNs) based on the author's choice(s), determine compliance for each matching journal, and present the author with a list of results.
3. An Institution could similarly determine compliance across a specified list of journals in a library collection, as covered by a TA with a publisher, or within a faculty area of discipline.
4. Publishers can run compliance analyses across their portfolio of journals, or offer a list of compliant journals for an institution under a TA.
5. Funders could specify a list of its preferred journals, or a list similar to those of institutions.

DATA SPECIFICATION

We used the generic structures described above as a framework for our detailed analysis. Having analysed the Plan S requirements line by line, and reviewed the various resources realistically available, we recommend the following specific approach.

CORE IDENTIFIERS AND COMPLIANCE ROUTES

For an author to determine compliant publishing options for one journal, each core identifier applies to one or more compliance routes as follows.

Compliance Route / Core identifier	FULLY OA	REPO-SITORY	TAs	TJs
Journal ID	✓	✓	✓	✓
Repository ID	✗	✓	✓	✗
Funder ID	✗	✗	✓	✗
Institution ID	✗	✗	✓	✗
TA (Agreement) ID & Date Range	✗	✗	✓	✗

CORE DATA SOURCES

All User Stories can be addressed by cross-referencing various lists of information, against which core identifiers can be matched. We recommend the following sources so we can

- build on the work already done, e.g. through DOAJ, RoMEO, etc.
- "separate concerns" – i.e. follow best practices of allowing each contributing component to focus on its specialty

Note that the sources may need to add fields or processes, as noted in the "Assessment of Sources" section below.

List	Proposed Source
1. A list mapping ISSNs to their variants (print and electronic ISSNs)	ISSN.org (albeit with caveats about data quality noted later).
2. A list of Fully OA (Fully OA) compliant journals	DOAJ
3. A list of Subscription/Repository route journals	RoMEO
4. A list of compliant Repositories	OpenDOAR
5. A list of TAs	ESAC TA Registry
6. A list of which journals exist within a given TA	We assume each Consortium (or paying institution) maintains its own list. **No central source exists**. However, if federated sources use consistent identifiers and offer APIs
7. A list of which institutions participate in a given TA (including their domain names)	

	(or download facilities in standard structured formats), then a central resource could be collated.
8. Which funders recognise that TA as 'transformative'	**No source exists.** (Raised in Open Questions, above.)

DATA ANALYSIS

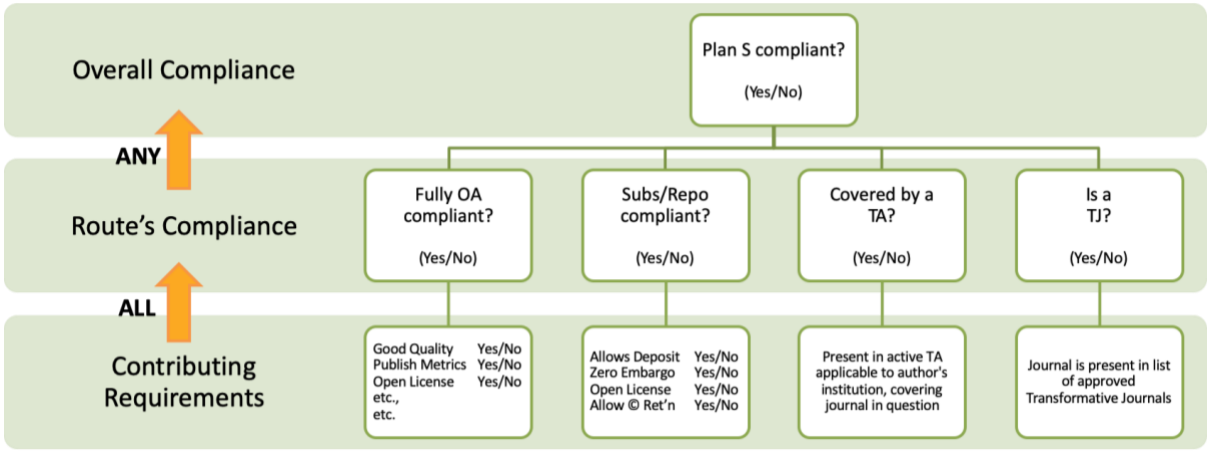
The bulk of the detailed analysis is supplied in the accompanying spreadsheet to this report, *JISC Plan S Data Spec.xlsx*. It contains several worksheets (tabs) as follows:

1. **Plan S Requirements**
    a. Analyses Plan S requirements line by line to determine data needed.
    b. Suggests fields to indicate **Contributing Requirements**
        i. A "**Yes/No**" for each requirement
        ii. **Optional** descriptive fields.
    c. Calls out detailed questions or assumptions for each requirement.
2. **Source Field Details**
    a. List of key sources and their native fields to analyse data available.
    b. Analyses which might be candidates for a Plan S Compliant data specification
3. **TAs and TJs**
    a. Specifies a bare minimum set of data to track TAs and TJs.
    b. Fills this key gap in specification as there are no generic sources available.
4. **Mapping**
    a. Maps data needed to data available (i.e. items 1 + 2 + 3).
    b. Basis for determining logic for compliance based in fields from key sources.
    c. Details gaps in data.
5. **Logic**
    a. Specifies how to combine the data in item 4 to determine compliance.
    b. Weaves together the hierarchy of data needed for each route.
    c. Focuses on mandatory requirements only; simplifies the details in the other tabs.
6. **Sources Assessment**
    a. Provides details of key sources assessment (see below)

LOGIC TO DETERMINE COMPLIANCE

Overall Compliance is determined by examining compliance for each route. Each **Route's Compliance** in turn is determined by examining its respective **Contributing Requirements**. We therefore have a hierarchy of data, as summarised below. The accompanying spreadsheet provides detail mappings to fields from key sources (DOAJ, OpenDOAR, etc.)

The figure below shows how the hierarchy of data forms our "compliance pathway."



- We have four data fields, one for each compliance route. If any one or more of the Route's Compliance fields are a "Yes" then the venue is deemed compliant overall.
- Each route's field would be set to "Yes" if the venue is compliant under the specific route. This can only happen if ALL the Contributing Requirements for that route are a "Yes."

For the Subscription/Repository route, the journal must have at least one license option that allows an open license and copyright retention, and which can be used in conjunction with deposit of articles. To avoid redundancy, we do not include fully open journals in Subscription/Repository compliance (even though they technically comply).

We envisage that the developer of a compliance tool would gather the specific field values for a given venue (or venues) from relevant sources, and then apply the logic as stated to determine compliance for each compliance route in turn.

A summary of the logic based on the fields analysed in the spreadsheet is as follows. We use a form of "pseudo code" which describes the data structure generically. Developers building tools should be able to translate this into details specific to their chosen development environments. Full details are provided in the accompanying spreadsheet.

OVERALL COMPLIANCE			
Overall Compliance =	JrnlIsFullyOACompliant = 'Yes'	Note: journal is compliant if	
	OR JrnlIsRepoCompliant = 'Yes'	ANY ONE or MORE of these	
	OR JrnlisUnderTA = 'Yes'	is "Yes"	
	OR TJJrnalIsCompliant = 'Yes'		
Route Compliance	**Contributing Requirement**	**General Notes**	**Gap**
FULLY OA			
JrnlIsFullyOACompliant =	JrnlHasQualityStandard = 'Yes'	(Same as Ed. Policies?)	
	AND JrnlHasEditorialPolicies = 'Yes'		
	AND JrnlHasPublishingMetrics = 'Yes'	No-one has these yet	YES
	AND JrnlAllowsCopyrightRetention = 'Yes'	(Implicit for fully OA jrnls)	
	AND JrnlHasOpenLicense = 'Yes'	(Implicit for fully OA jrnls)	
	AND JrnlAllowsDeposit = 'Yes'	(Implicit for fully OA jrnls)	
	AND JrnlEmbargoLength = 0	(Implicit for fully OA jrnls)	
	AND JrnlIsInDOAJ = 'Yes'	+ other criteria	
	AND JrnlHasPricingInfo = 'Yes'		
	AND JrnlHasCostsInfo = 'Yes'	No-one has these yet	YES
	AND JrnlAllowsWaivers = 'Yes'	Only applies if APC charged	
	AND JrnlHasID = 'Yes'	Need ISSN-L to ISSN(s)	YES
	AND JrnlHasArticlePIDs = 'Yes'		
	AND JrnlHasPreservation = 'Yes'		
	AND JrnlHasCoreArticleMetadata = 'Yes'	Definition unclear	YES
	AND JrnlHasArticleLicenseInfo = 'Yes'		
SUBSCRIPTION/REPOSITORY			
JrnlIsRepoCompliant =	(JrnlAllowsDeposit = 'Yes'		
	AND JrnlEmbargoLength = 0		
	AND JrnlAllowsCopyrightRetention = 'Yes'		
	AND JrnlHasOpenLicense = 'Yes')	Journal must have an option	
RepoIsCompliant =	RepoIsInOpenDOAR = 'Yes'	Has an OpenDOAR ID	
	AND RepoHasArticlePIDs = 'Yes'	OpenDOAR does not track	YES
	AND RepoHasArticleMetadata = 'Yes'		
	AND RepoHasArticleOAStatus = 'Yes'	OpenDOAR does not track	YES
	AND RepoHasAvailability = 'Yes'	OpenDOAR does not track	YES
TRANSFORMATIVE AGREEMENT			
JrnlisUnderTA =	JournalID IN list of applicable ISSNs, determined as follows:		
	1. Look up Author's Institution_Domain in TA_Participants table to find Consortium_ID		
	2. Look up Consortium_ID in TA_Consortia table to find TA_ID(s) for institution		
	3. Look up Funder_ID in TA_Funders table to find TA_ID(s) for funder		
	4. Look each TA_ID from steps 2 and 3 up in TA_Journals table to infer…		
	…a list of ISSN_L(s) covered by applicable TA(s)		
TRANSFORMATIVE JOURNALS			
TJJrnalIsCompliant =	JournalID IN TJ_Registry.ISSN_L	**Interim solution if a registry used**	
	- OR -	(When checks fully implemented)	
TJJrnalIsCompliant =	JrnlHasOAuptakeInfo	No sources exist.	YES
	AND JrnlHasOAuptakeInfo	(Use RoMEO?)	YES
	AND TJOAcontentAvailable	No sources exist.	YES
	AND JrnlHasOffsetting	No sources exist.	YES
	AND JrnlIsCostNeutral	No sources exist.	YES
	AND JrnlHasAuthorMetrics	No sources exist.	YES
	AND JrnlHasAnnualReport	No sources exist.	YES

HANDLING AMBIGUITIES

Where gaps exist, we discussed a phased approach with the Project Steering Group.

- Journals are assumed compliant (given a "temporary free pass") in ambiguous areas until the ambiguities are resolved.
- cOAlition S follows a phased approach, so ambiguous criteria are brought in over time as they become fully defined, with compliance with those criteria being assumed during the interim period of their being unknown.

cOAlition S members will need to decide priorities and timing for the phasing. The key ambiguities or gaps are detailed in the "Open Questions" section above. In summary they are:

1. No industry-standards or sources exist for information about publishing statistics.
2. No industry-standards or sources exist for information about publishing prices and costs. The Plan S requirements are unclear about exactly what information is required.
3. Requirements specifying "in the process of being registered" in the DOAJ or OpenDOAR would not be workable in practice. The sources do not implement this process, and how would rejections be handled?
4. For requirements stating "at no additional cost," how is this calculated or defined?
5. Formats for metadata are not specified for several requirements where PIDs, "quality metadata" and "machine readable" metadata are mentioned.)

ASSESSMENT OF SOURCES

DEFINING "QUALITY"

The brief specified assessing criteria such as reliability, validity, terms of use, etc., focusing on what could reasonably be implemented by December 2020. We will define "quality" as "fitness for purpose", to include the factors above.

Many factors are a matter of judgement or may not be possible to quantify. For example, there is no definitive list of journals, so we have no control group against which to quantify a particular source's journal coverage. During discussions with the Project Steering Group, we agreed that a simple "High/Medium/Low" structured qualitative analysis would be sufficient to provide actionable information.

QUALITY MATRIX

We assessed key sources according to the following criteria.

Criterion	Definition	Questions/Notes
Coverage	Per-article/per-journal/per-etc. (Description and # Journals)	Focus on data needed for User Story.
Data granularity	Field coverage & taxonomies. ("H/M/L")	Is data sufficient & well-structured to indicate compliance?
Reliability	Rigour of technical operation. ("H/M/L")	Backups, scale-ability, development pipeline.
Validity	Accuracy of data. ("H/M/L")	% coverage; editorial checking/validation
Currency	Frequency of updates. Pro-active or via submission.	Important for self-declared Whitelists.
Terms of use	Data license – CC xx where possible.	Note other terms of reuse.
Sustainability	Funding & owners. ("H/M/L")	# years' funding; governance; # FTEs.
Authority	Perceived quality. ("H/M/L")	Anecdotal evidence and industry knowledge.
Links to other data	Dependencies.	We will look for potential links too, e.g. ISSNs & ISSN-Ls.
Legal issues (e.g. GDPR)	Personal data.	Awareness of issues; process in place
ISSN/PID Use	Are journal entries identifiable by ISSN	Pre-requisite for cross referencing of information
OA Status	Indicator of Fully OA/Hybrid/etc.	Working towards an index of esp. Hybrid & non-OA
API	Yes/No	Important for integration with other sources

Bulk download	Yes/No	Important for integration with other sources

SOURCE ASSESSMENT

The bulk of the assessment has been undertaken, and the results captured, in the spreadsheet accompanying this report. We summarise the various data sources and overall conclusions below.

The following tables summarise our assessment of different sources. We split sources into three groups:

1. **Core Sources** provide the bulk of the compliance data and are readily extensible
2. **Supplementary sources:** that may be used to determine and feed in supplementary information, or information that tools would need in practice.
3. **Other sources:** may be included in sources above, or not directly relevant, but which will crop up during discussions
4. **Sources Not Available**: how we might address data required, but not currently available in centralised, industry standard databases

We examined data points in detail and surveyed providers of major data sources, which are marked "Y" (for "Yes") under the Full Assessment column. Those marked "N" ("No") were assessed via desk research. Further details, analysing key sources against each of the quality criteria above, are captured in the spreadsheet accompanying this report (in the "Sources Assessment" sheet). We summarise the various data sources and overall conclusions below. The table key as used below is: ✓ - Applicable (i.e. the source may cover a **subset** of data needed); ✗ - Not applicable; Y – Yes; N – No.

CORE SOURCES

Core sources nominated by cOAlition S are as follows.

Source	Summary of Scope	FULLY OA	REPOSITORY	TAs	Full Analysis	Assessment Summary
DOAJ	Fully OA Journal policies	✓	✗	✗	Y	With incremental work, and under the assumption that cOAlition S will trust their editorial judgement, could provide complete data for Fully OA route. Robust infrastructure and editorial processes. Use placeholders for Costs & Journal Metrics info.

RoMEO	Journal policies	✓	✓	✗	Y	Could provide comprehensive coverage of Subscription/Repository-compliant journals. Robust infrastructure and comprehensive coverage. On-going concerns raised about responsiveness, and many publisher records appear dated. Data are not CC0. Could also be leveraged to add a TJ compliant flag?
OpenDOAR	Repositories	✗	✓	✗	Y	Does not track the data needed to determine of Repositories are compliant. A white-listing system would need to be implemented (as per the February 2019 workshop), with levels of curation to be agreed with cOAlition S.
ESAC	TA tracker	✗	✗	✓	Y	Has TA registry; would need to add field for compliant journals. Currently does NOT track the mapping between specific institutions and journals for each agreement.

SUPPLEMENTARY SOURCES

The following sources could be useful over time for building out tools.

Source	**Summary of Scope**	**Full Analysis**	**Assessment Summary**
Sherpa JULIET	Funders policies	Y	Not directly applicable here.
Sherpa FACT	Funder Compliance	Y	Check (UK) funder/author compliance – an example of an author compliance tool.
Crossref	Journal list	Y	Could form basis of global journal list.
Unpaywall	Article licenses	Y	Based on CrossRef DOI list; adds more info on journal type.
issn.org	List of ISSNs & ISSN-Ls	Y	Key underlying data source to handle multiple ISSNs per journal. However, note ISSN-L coverage is not extensive but not complete. E.g. DOAJ have found it necessary to regularly validate all ISSNs themselves.

OTHER SOURCES

Possible suppliers or partners – these are sources we were asked to consider by the Project Steering Group during discussions. We spoke directly to the following sources.

Source	Summary of Scope	Full Analysis	Assessment Summary
QOAM	Qualitative assessments of journals	Y	Would have a lot of ground to make up to populate the required data. Tracks users' perceptions of journal quality, but data does not currently include Plan S-specific journal metrics (rejection rates, etc. – although they are looking at Fair Open Access Alliance guidelines). Data include an initial "Plan S" flag, for now only covering the 4 mandatory technical requirements. Does not assess the journal. Data can accommodate mapping between agreements and journal lists for offsetting deals. Data is open under CC0. Position themselves as a vendor to host data. They are currently seeking funding (from CWTS). Minimal curation - a crowd-sourcing model. They have suggested that checking of random samples could be implemented if they were to scale, but this is not currently in place.
OA Switchboard	Linking hub of key metadata.	Y	Working towards a pilot through 2020. Data consumed is planned to cover journals' OA policies eligibility against funder requirements and payment requests.

For background, we looked at the follow sources via desk research

Source	Summary of Scope	Assessment Summary
Norwegian Register	Classified journal list	Information on journal classification across ~30k journals. An example of a local list used in practice.
Delta Think OA DAT	Combines lists for journal types & APCs	Shows principles of central journal lists and tracking pricing information. Focus on large publishers – would need to be extended to cover long tail. Closed now, but could be made open if funded. (**Note:** Delta Think is producing this report.)
Bielefeld Gold OA List	Fully OA journal list	Extended list of known, but uncertified Fully OA journals. Illustrates practice of using ISSN-L to collate multiple sources. They would need to add Plan S certification information.
GOAJ	OA adoption & patterns	Data from librarian Walt Crawford data. Mirrors a subset of DOAJ data.
SciELO	S. American journals	Source of ISSNs, e.g. to supplement CrossRef, but would need DOAJ checks adding. Might help overcome the technical hurdles of Plan S for no-fee/diamond journals - would need to coordinate with DOAJ to analyse the overlap between the two sources.

SOURCES NOT AVAILABLE

Finally, we know that some data required by Plan S are not currently available in centralised, industry standard databases. The following table identifies data sources that would need to be assembled in order to fill in the key gaps.

Summary of Scope	Notes
List of TA details: participating journals and institutions	Central database needed; currently locally maintained by consortia (and sometimes publishers.) We specify an outline of the data structure needed in the accompanying spreadsheet's "TAs and TJs" sheet.
Database of Transformative Journals	Central database needed. Could this build on Sherpa RoMEO data? We specify an outline of the data structure needed in the accompanying spreadsheet's "TAs and TJs" sheet.
Industry-standard database of Journal Metrics	To cover at least Plan S-compliant journals.
Database of costs	No definitions agreed by cOAlition S, pending outcomes of Information Power study published during the week the final version of this report was submitted. Assume a placeholder flag for now in DOAJ.
Database of list prices	Unclear if it is needed. A mix of Delta Think and DOAJ data could form a starting point. (**Note:** Delta Think is producing this report.)

It is worth noting that no single database to cover all compliance routes is being discussed. The approach to measuring Plan S compliance is essentially one of multiple whitelisting. Where journals are not known about, non-compliance is inferred (e.g. if a fully OA journal is not in DOAJ, it can't be compliant with the Fully OA route).

CONCLUSIONS AND RECOMMENDATIONS

During the project, the Project Steering Group indicated that they wanted to seek our opinions. The following conclusions therefore represent a combination of the structured assessments undertaken, and our views based on anecdotal feedback and our industry knowledge.

cOAlition S has made clear its need for speed of implementation, so we have therefore prioritised the most mature, well-scaled providers in our assessment.

CHALLENGES

Each compliance route has different needs. The approach to measuring Plan S compliance through whitelisting of a few key sources is, in our view, the most practical approach. Note that there is a difference between **curation** (ensuring data is accurate) and **collation** (making curated data available via a convenient source). Ideally, data sources must offer both. The biggest challenge to them is curating the data across the thousands of entries they cover. The key sources selected (DOAJ, OpenDOAR and RoMEO) are mature and robust. They are good starting points, but do not (yet) cover all the requirements. ESAC does not curate its data – it (currently) only acts as a collation point. We provide a review for each compliance route as follows.

Compliance Route	Comments
Full OA	The DOAJ covers most data needed and could realistically be extended to cover the rest during 2020. The team are currently trying to plan and pre-empt requirements, but have indicated that technical development requires funding to be put in place. Our key concerns here are: • There may be additional level of rigour needed for Plan S checks. During discussions we clarified that the DOAJ editorial team's judgement is deemed sufficient. • The DOAJ cannot certify ambiguous requirements. Hence, we recommend the phased approach outlined in the "Prioritisation of Requirements" section above. • The DOAJ does not support the notion of "in the process of being registered" as we discussed in the "Handling Ambiguities" section above. We suggest temporary resources should be made available to support a one-off surge of registrations. • The DOAJ relies on publishers proactively depositing data. Data compliance is out of scope of this study, but some process of checks between publishers and the DOAJ would need to be implemented ensure data is deposited and maintained to standards and timeliness acceptable to cOAlition S.

	• The DOAJ does not appear to have a unique journal identifier, but instead relies on a combination of print and electronic ISSNs. Its data is clean so this is a minor issue.
Subscription/ Repository	Sherpa RoMEO has the data needed to determine if **Journals** allow deposit. We were not able to obtain formal technical details (e.g. field names), but could analyse data based on the user-facing results (in the newer version [Version 2] of the UI). Our key concerns here are: • A key challenge lies in unpicking multiple options where journals offer a choice (e.g. zero embargo deposit under restrictive license vs. embargoed deposit under open license). The team has experience of handling this and delivered tools such as Sherpa FACT, which demonstrate the principles of compliance checking. They also have experience in resolving nuanced details, such as ambiguities around split journal ownership • The data are not offered under a CC0 license. Our understanding is that this is to avoid concerns about "free riders" and allow the team to focus its resources on the needs of its direct sponsors. **cOAlition S would need to resolve this with Jisc.** • Responsiveness to publishers. We have heard from multiple sources that they perceive the Sherpa team to be unresponsive – even after the team initiate requests for data. We found data records showing updates from months or years previously that publishers told us were out of date. The process of depositing data with Sherpa is seen as being harder than working with the DOAJ, even though in practice it appears to require similar levels of effort. **The Sherpa team's intent and infrastructure are trusted. The perception is that issues are due to lack of resources.** • RoMEO relies on publishers proactively depositing data. Data compliance is out of scope of this study, but some process of checks between publishers and RoMEO would need to be implemented ensure data is deposited and maintained to standards and timeliness acceptable to cOAlition S. OpenDOAR does **not** currently track the data needed to determine of **Repositories** are compliant. However, it remains the only comprehensive database of content repositories available as a realistic starting point. We think a white-listing system (as suggested at the February 2019 workshop) would appear the most practical way of implementing something quickly, but resources would then be needed to bring detailed data and checks up to speed. For now, we assume the presence or absence of a repository in OpenDOAR implies compliance.
TAs	We assume ESAC could add a flag to their data set to indicate if a given TA is compliant. They have indicated that this is possible, but

	they would require someone else (the cOAlition S project office?) to curate the flag and determine a given TA's compliance. **Further, a new database would need to be built to centrally collate details about which specific journals are compliant for which institutions.** This further level of detail lies outside ESAC's current overage. Implementing the new database would require the commissioning of an extension to ESAC or the use of a 3rd party provider. (Again, ESAC would not curate the data.) The spreadsheet accompanying this report provides a design for the structure of a TA database. Discussions during the course of the projected suggested that individual consortia would be responsible for curtaining (maintaining) the data specific to them. For offsetting deals, which have the same contract structure as TAs, QOAM's data structure can accommodate institution-specific information for every contract.
TJs	**A new database would need to be built to centrally collate whether a given journal is deemed compliant**.

Other sources either cover a much smaller subset of data than the ones noted above, or are pitching themselves as collators but not curators. Some are simply less well-scaled or mature than the core sources nominated by cOAlition S. Some concerns were raised to us that, in nominating particular sources, cOAlition S is creating a "monopoly" in certain areas. However, on balance, we do not consider this to be an issue. For each route it is important that one source only is deemed to have authority and offer a "single version of the truth" so robust and effective assessment can be actioned, and conflicts between different data sources can be avoided. We anticipate that a responsible authority would organise appropriate feedback and community input to ensure that its data is accurate and fair. In principle, any properly structured data collated could be transferred to alternative providers in future.

RECOMMENDATIONS

In our view, the next steps and priorities should be as follows.

FOR ALL COMPLIANCE ROUTES

1. Agree on phased approach and take clear decisions on the requirement priorities.
    a. These are detailed in the "General Guiding decisions on determining compliance" section above.
    b. The data specification is flexible, allowing for indicators of presence or absence of required data or features. The indicators can be defaulted to "compliant" in the (interim) absence of hard data.
    c. We suggest cOAlition S runs a small project to produce a rough timeline for implementing de-prioritised requirements, so all stakeholders can manage expectations and work towards solutions over the longer term.
    d. We suggest that priority should be given to specifying a limited taxonomy for license information embedded in articles. It is a key measure.
2. Clarify the delegation of the curation of data for each route to a specific authority. ("Authority" is taken here to mean "trusted provider.") This is already in progress, but we suggest quickly finalising details as time is tight for 2020 implementation:
    a. Work with each Authority to agree the most important requirements needed, where gaps exist, scope out the work needed to fill them, and ensure budgets are in place.
    b. Agree formal guidelines that empower the Authority to take editorial decisions. The DOAJ raised this (see "Authority of Trusted Sources" above), so issuing some guidance may help all concerned.
    c. Establish agreement between Authorities about which owns which master data. E.g. DOAJ information on fully OA compliance should always be considered the primary source, and RoMEO data should reflect DOAJ if RoMEO was to add a "Fully OA Plan S compliant" flag to its data set.
3. Commission some work to flesh out general data ownership and maintenance principles. We suggest working with the authorities nominated, AND running a focus group with a few key publisher representatives to tease out workability issues. (This may naturally follow from the project plan to implement the author tool, which is running in parallel to this investigation.) As starting point for discussion, we would suggest:
    a. Each Authority (data source) operates on a whitelisting principle. It should only list venues that may be compliant for each compliance route, and assume that any not listed are not compliant for that route.
    b. Agree with each Authority its approach to maintaining data quality. Editorial policies may vary, and represent a trade-off between cost and accuracy. There is a balance to be struck between self-certification (cheap, but at the mercy of 3rd parties) and entirely independent curation (expensive, but highly accurate).
    c. Another point of discussion is whether publishers should be formally mandated to keep authorities up to date, and what "service levels" (levels of responsiveness) authorities should offer in return.

d. We suggest some sort of proactive deposit of compliance metadata plus random spot-checking policy would strike the right balance between cost and accuracy.
e. Authorities will also need to form a view on minimum acceptable timeliness of data (e.g. whether it should be updated weekly/monthly/quarterly).
f. We recommend running a focus group involving whitelist operators and a few key publishers to clarify the best balance between voluntary or mandated deposition of compliance data, responsiveness and rigour of data validation. The results could be used to set expectations and foster understanding between all stakeholders.

FOR SPECIFIC COMPLIANCE ROUTES

Route	Recommendations
Fully OA	The DOAJ covers most of the data needed and could realistically be extended to cover the rest during 2020: • Formalise decisions with the team, so their remit is clear. Discussions during this project clarified that their editorial judgement and process are acceptable. • Funding needs to be made available in a timely way. • Clarify the dropping or mitigation of the "in the process of being registered" requirement, per discussions. Undertake a small study to quantify anticipated spike in demand. • Clarify priorities of requirements and how to handle any that have been de-prioritised. • The DOAJ does not appear to have a unique journal identifier, but instead relies on a combination of regularly validated print and electronic ISSNs. Tool developers may wish to work with them to develop a DOAJ ID.
Subscription/ Repository	Work with Sherpa team/Jisc to • Formalise the use of Sherpa RoMEO to indicate whether journals allow deposit. • Agree a CC0 license for relevant RoMEO and OpenDOAR data. • Address concerns about responsiveness to publishers' data deposition. (We think this is likely to be about agreeing prioritisation of resources.) • Tap into their expertise to help cOAlition S resolve the challenges of ambiguities, such as multiple authorship or multiple journal ownership. • Similarly to the DOAJ, agree editorial ownership, requirements priorities (particularly around Repositories).
TAs	TAs represent a particular challenge due to their complexity. We do not think it will be possible to appoint one single "authority" to **curate** the detailed data. • cOAlition S should determine a clear policy on who is responsible for maintaining and curating the list of journals and institutions for a given TA. (Note that these lists are fluid, and may change through the duration of a TA.) The steering group suggested this should be the buying

	consortium, which may in turn specify contractual service-level obligations to the publisher. Funders should maintain their own lists of journals. The example from the Netherlands' SURFmarket implementation may serve as a good case study. • cOAlition S should work with ESAC to add a "Plan S compliance" indicator to ESAC, and clarify curation responsibilities as above. • cOAlition S should identify a provider, and agree and fund further work to build and maintain central machine-readable database of which journals apply to which agreements. (This currently lies significantly out of scope of ESAC's current coverage.) ESAC or the Netherland's SURFmarket are logical places to start. Individual consortia would be held responsible for populating the central database. • Holding a workshop or focus group involving some key publishers and consortia may prove useful to tease out operational details. The spreadsheet accompanying this report provides a draft design for the database structure of a TA database. The Netherlands' SURFmarket provides a specific case study of a workable implementation. Use its experience to help tease out details of how to handle policy exceptions.
TJs	Facilities to identify and track TJs need to be built. • cOAlition S should determine a clear policy on who is responsible for maintaining and curating the list of approved journals. • cOAlition S should identify a provider, and agree and fund further work to build and maintain a central machine-readable database of approved journals.

DECISIONS FOR COALITION S AND TOOL DEVELOPERS

The Compliance Task Force asked us to comment on decision-making priorities regarding the data specification. Our recommendations are as follows.

We recommend that **cOAlition S decides** on the following before inviting tenders for the compliance checking tool:

- which mandatory requirements are needed for launch;
- policy details about mandating compliance data deposition (or not) and data verification;
- who is responsible for curating data for each compliance route (and agreeing budgets and expectations with them);
- rules for multi-author papers and specifying policy exceptions.

Timing is already tight for 2020 implementation, so we recommend cOAlition S quickly agrees budgets and expectations with the key sources responsible for curating data for each compliance route (e.g. DOAJ, Sherpa, ESAC), so they can proceed with any necessary implementation.

We anticipate that the following details would be handled by the **tool's developer**:

- the process for escalating and resolving questions about the data;
- details of engagement with data providers, end users and publishers (if applicable);
- data update frequency and processes;
- specific metadata taxonomies.

APPENDIX – PROJECT TERMS OF REFERENCE

For reference, we outline the scope of the project as follows.

As Plan S is implemented, the following four **user stories** become important (the first one is the priority – the rest probably follow):

a. Authors need to know quickly, easily and clearly what their compliant† publishing options are – this is the tool noted above
b. Institutions need to be aware of what researchers' compliant publishing options are, so they can advise on how those options align with any institutional policies.
c. Publishers need to be aware what researchers' compliant publishing options are, so that they can fill gaps where these might exist
d. Funders need to be aware what researchers' compliant publishing options are, so they can monitor the progress of Plan S, provide a tool to support (a) and, where necessary and appropriate, take measures to fill gaps.

Objectives of the project:

1. Draft a specification for the data needed to meet the four user stories outlined above, for all three Plan S routes to open access, including the data type, level of granularity, currency, reliability, authority and links to other data. If some prioritisation is needed within the review, then the focus should be on user story (a); the tool for researchers.
2. Review possible sources of the data in the specification, based on current provision and development work that can reasonably be expected to be completed by April 2020, given modest investment. The review should cover whether the data are held, their reliability, validity, currency, terms of use, sustainability, legal issues (eg GDPR), and any other relevant factor.
3. Recommend the best data sources.
4. Identify medium term (one year) strategies to fill, or mitigate for, gaps or other shortcomings in the data, and estimate the associated risks and costs. One strategy may be the use of "white lists", or managed self-declaration by journals or platforms that they are compliant.
5. Validate the findings with a small number of key experts including those working on the data sources concerned and members of the relevant cOAlition S task force.

† - "compliant" here means compliant with the Plan S principles and implementation guidelines released 31 May 2019. It is recognised that individual members of cOAlition S may adopt policies that have variations on these guidelines, but those are out of scope for this project.