DNA3.1 - Report on the coordination of accounting data sharing amongst Infrastructures (initial phase)

Publication Date	2018-04-30
Authors:	Uros Stevanovic;David Groep;Ian Neilson;Stefan Paetow;Wolfgang Pempe
Document Code:	**AARC2-DNA3.1**
Contractual Date:	30-apr-2018
Actual Date:	30-apr-2018
Grant Agreement No.:	730941
Work Package:	NA3
Task Item:	TNA3.2 “Service-Centric Policy”
Lead Partner:	KIT

The research leading to these results has received funding from the European Community’s Horizon2020 Programme under Grant Agreement No. 730941 (AARC2).

Abstract

This report presents the results of the desk study on the evaluation of risks to (personal) data protection as considered in the European General Data Protection Regulation (GDPR), for Infrastructures and their service providers that leverage federated identity management (FIM) to connect research and collaboration users. Specifically, it considers personal data collected as a result of using the infrastructure (not any risks relating to the research data itself, which is a community responsibility) and provides guidance to the Infrastructures concerning Data Protection Impact Assessment (DPIA) in the FIM context. The authors present recommendations to Research Communities for determining the necessity of formal DPIA and guidelines for its execution.

This report does not constitute legal advice in any specific jurisdiction.

Table of Contents

1. Introduction

The Research and e-Infrastructures, in the course of their regular activities of providing services to the research community, will inevitably collect a variety of personal data: records of access to a compute service; the name (in whatever form) of the initiator of a transfer of personal data; a history and archive of past service usage to support accounting (and for determining exhaustion of any resource allocations to the user or research community); for justification of their own usage of resources if the service or infrastructure is open i.e.does not rely on 'pay per use'. The majority of the data so collected by the infrastructure can – in a direct or indirect way – be linked to a person, and thus falls within the scope of the data protection regulations in Europe.

The use of federated identity management (FIM) in itself already goes a long way to satisfying key principles of the General Data Protection Regulation (GDPR) [GDPR] e.g.: data minimisation - personal data provided about the user during authentication is already limited to necessary elements (attributes); data security - passwords never leave the identity provider of the user's Home Organization. The "IdP-SP-Proxies" of the Blueprint Architecture allow limiting the incoming attributes to the "research and scholarship" set, which is basically the same set of attributes as would be released by the researcher simply sending an email. Yet still many FIM identity provider organisations are hesitant to release attributes because of a perceived liability when releasing even a name, a unique identifier, and an email address for their researchers. The work on the GÉANT Data Protection Code of Conduct ("DPCoCo") [DPCoCo] in the REFEDS community addresses this issue by allowing service providers and Infrastructure proxies to confirm explicit adherence to GDPR principles, thereby allowing relying FIM participants to have more trust in the service or infrastructure. Although elements of this framework still need to be resolved, it would – when adopted and appropriately endorsed by the European Data Protection Board – go a long way to making necessary attributes available to the Infrastructures and proxies.

The DPCoCo, whilst acknowledging accounting as a requisite part of granting access to services and bringing community-managed personal attributes of users within its scope, does not address the fact that personal data collected by the infrastructure as part of its operation also results in 'new' personal data such as the association of the workflow usage data, network identifiers (IP addresses) or specific dataset identifiers to the person. In an ecosystem where infrastructures are interconnected (and where research infrastructures use each other's as well as generic e-Infrastructure services from a catalogue in a dynamic way), such data may be shared between many different parties, who are all independent data controllers – each of them by itself determines both purpose and means of the processing.

The most visible personal data collected by the Infrastructures as part of their operations is accounting data. Indeed, for managing cross-infrastructure resource allocations, the sharing of accounting data is necessary for enabling access to federated services in the ecosystem, as well as being 'novel' in the sense that correlated accounting data may reveal workflow usage and research operating patterns of the individual researchers.

Sharing of accounting data between generic federated e-Infrastructures and homogeneous communities was studied and described in the (AARC1) guideline G016 [AARC-G016]

Recommendations on the exchange of personal data in accounting data sharing. That guideline deals with generic accounting information exchange and does not address intra-community needs for accounting, e.g. based on community-specific authorisation attributes such as groups and roles. Extending these recommendations on the protection of personal data in sharing accounting data to more complex communities (with significant internal structure and with internal controls) requires two elements: *(i)* understanding the risks resulting from the processing of Infrastructure-generated personal data within a community, and *(ii)* understanding the extent to which the organisational structure of a (research) community itself has to be reflected in the exchange of (accounting) data generated by the Infrastructure (because distinct groups exist within the community and there is no single community manager responsible for all of the data).

n this initial phase recommendation, we focus on the risk assessment for personal data processing for collaborative and research communities in the context of the GDPR. The GDPR recognises the concept of a Data Protection Impact Assessment (DPIA) which comprises both an assessment framework to determine whether a processing is likely to result in a high risk for individuals, as well as a methodology to assess what the impact could be if there is indeed such a high risk. However, the guidance from the European Data Protection Board (EDPB, until recently “WP29”) must be interpreted in the context of research and collaboration infrastructures that are already using FIM mechanisms (reducing the risk), have adopted the Blueprint Architecture IdP-SP-Proxy concept (which harmonises the attribute management, but encourages cross-domain services involving many data controllers and use of omnidirectional identifiers for individuals), and exchange (accounting) information that has been collected in many services as a result of the user’s actions. The resulting risk assessment, likely having many common elements between the Infrastructures, can thus be made easier for the research communities if the DPIA guidance is targeted specifically to research communities employing FIM.

At the same time, it should be recognised that some of the basic principles highlighted in Opinion 248 (rev. 01) from the EDPB, even when the processing results in a high risk, may already have been addressed once the DPCoCo has been endorsed. Similarly, seeking the view of the data subjects (researchers) is a feasible route, especially for well-coordinated communities (keeping in mind that seeking a data subject’s viewpoint cannot be done by asking for consent at the time of processing). So even when a DPIA is necessary for the processing (which in itself will depend much on the community involved), there may be sufficient commonality in the FIM Infrastructures and in the IdP-SP-Proxy operations to lighten the burden on the communities and Infrastructures.

In this document, we put the above considerations into context by giving background on the current GDPR structure and its prevalent interpretation by the EDPB and the R&E data protection experts (including DPCoCo and the GÉANT Task Force on data protection), discuss the methodology of the Data Protection Impact Assessment, and give guidance within the context of federated infrastructures for collaborative research communities and Blueprint IdP-SP-Proxy operators on how to determine whether a DPIA is needed, and if so how to perform such an impact assessment.

2. Data Protection Regime and the GDPR

The General Data Protection Regulation "2016/679" (GDPR) was adopted on 27 April 2016 and will come into effect on 25 May 2018. It has a profound impact, not only on society in general, but also specifically on international research and collaboration, where its global nature and the need for open science and sharing of resources present specific challenges. As stated in Article 3(1), the Regulation "applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not." It not only applies to entities that are established in the EU, but also to organisations located outside of EU if they "offer goods or services" or "monitor the behaviour" of data subjects residing in the EU – and as a Regulation it has direct effect, requiring neither further ratification or implementation into national law. And Infrastructure Services, by definition, offer services to both European and global researchers – and collect personal data in the process of offering these services.

The paper by Christopher Kuner entitled "International Organizations and the EU General Data Protection Regulation" [KUNER2018], presents is clear and short summary of the relevant GDPR provisions, with the main points logically grouped in the following way:

1. *The principle of lawful processing*: Recital 39 and Articles 5(1) and 6 are outline the legal basis for processing of personal data.
2. *The purpose specifications and limitation*: Article 5(1) states that the purpose of processing must be visibly defined before it is started.
3. *Data quality*: Article 5(1) states that the data must be removed if the processing is no longer taking place. Additionally, it outlines principles for processing like data minimisation, accuracy, storage limitation, integrity, and confidentiality.
4. *Fair processing*: Data subjects must be informed before their data are being processed about the purpose of the processing and about the identity of the controller. These and further rights of the data subjects are outlined in the Articles 12-22
5. *Accountability*: In the Article 5(2) it is stated that active security and privacy measures for the protection of personal data must be implemented by the data controllers. Controllers are responsible for the compliance of the processing operations with the data protection law. Also, compliance with the provisions of the law should be demonstrable by the data controllers to the data subjects, data protection authorities, and general public.

For the monitoring of the GDPR compliance and implementation the European Data Protection Board (EDPB), comprised of the national regulatory authorities of the member states, will be established on May 25th, 2018.

3. Data Protection Impact Assessment – DPIA

The GDPR, much more than previous directives and legislation, places the data subject in a central position, and organisations processing personal data have to continuously consider the effect their actions and processing has on the people involved. A key mechanism in the GDPR is “risk assessment”: the “Data Protection Impact Assessment” (DPIA) of Article 35. It is a favoured mechanism, and is also present e.g. in Directive 2016/680 on crime and prosecution data. Article 35 of the GDPR states that when the processing of personal data is “likely to result in high risk to the rights and freedoms of natural persons”, the data controller must conduct an assessment of the impact of the previously envisaged processing on the protection of personal data. So while not always necessary, at least a basic risk assessment it needed, also in research and collaboration infrastructures, and here specifically for Infrastructure-generated data.

This chapter presents the viewpoint of “WP29” (the future EDPB) expressed through its Opinions regarding DPIA, and the necessary steps and conditions involved with conducting a DPIA, i.e. Opinion WP 248 [WP29-248]. From the Opinion:

> *A DPIA is a process designed to describe the processing, assess its necessity and proportionality and help manage the risks to the rights and freedoms of natural persons resulting from the processing of personal data by assessing them and determining the measures to address them. DPIAs are important tools for accountability, as they help controllers not only to comply with requirements of the GDPR, but also to demonstrate that appropriate measures have been taken to ensure compliance with the Regulation (see also article 24). In other words, a DPIA is a process for building and demonstrating compliance.*

As stated, conducting DPIA may be necessary to show that the processing of the personal data, and measures taken for such processing, are compliant with the GDPR. Also, if DPIA was not conducted when it should have been, or was conducted improperly, this may lead to monetary fines, in this case up to 10 M€, or up to 2% of the global yearly turnover, whichever is higher. However, It is not always mandatory to conduct DPIA for every processing operation: it is required only when the processing is “likely to result in high risk to the rights and freedoms of natural persons”. But regardless, the obligation to “appropriately manage the risks for the rights and freedoms” remains. Effectively, this means that the risks, at the very least, need to be identified, analyzed, and evaluated. The risks should also be “reviewed regularly”.

It is expected that the EDPB will issue guidelines, recommendations, and best practices in order to have a consistent approach and application of the GDPR. Furthermore, EDPB, in the future, should issue further clarifications, including examples, how to conduct DPIA and for which processing operations, and whether the relevant supervisory authority should be consulted. Similarly, jurisprudence will have to be developed. Yet only little of such guidance is currently available that is of direct relevant to research and collaboration infrastructures.

The following schematic, taken from the WP29 Opinion referred to above, serves to indicate the basic decision process related to the DPIA in the context of the GDPR.

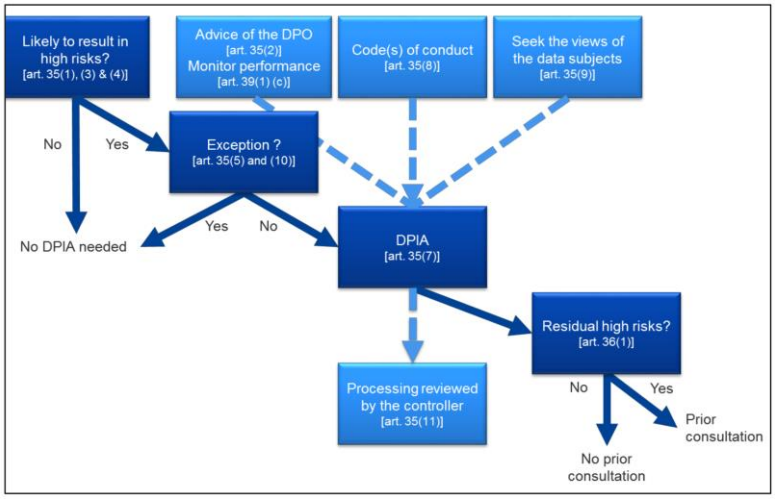


Figure 1: DPIA assessment flow diagram from teh WP29 Opinion 248 rev 1

The first stage in deciding whether to conduct a DPIA is identifying the risks. If there are no high risks (and one can reasonably substantiate it), the organisation (so the Infrastructure, FIM Proxy, or Community) is done.

If the risks are "likely to be high", and none of exceptions listed in Articles 35 apply (they don't for our purposes, unless the EDBP or member state law were to grant an explicit exception to Infrastructures!), then a DPIA must be conducted. In assessing the risks:

- any existing public availability of the personal information should be considered.
- in the process of conducting a DPIA, the DPO of the organisation conducting the assessment, if one exists, must be consulted. The view of the DPO may be ignored, but the reasons for such a decision must be documented.
- the existence of and compliance with a Code of Conduct must be taken into account when assessing the impact of identified risks.
- data subjects whose personal data are processed should be consulted "where appropriate". However, the consent of data subjects for processing is not a way to seek their opinion on it (i.e. one cannot 'abuse' a consent button as a way of of claiming that 'the user was consulted' and using that in the assessment to claim a lower risk).

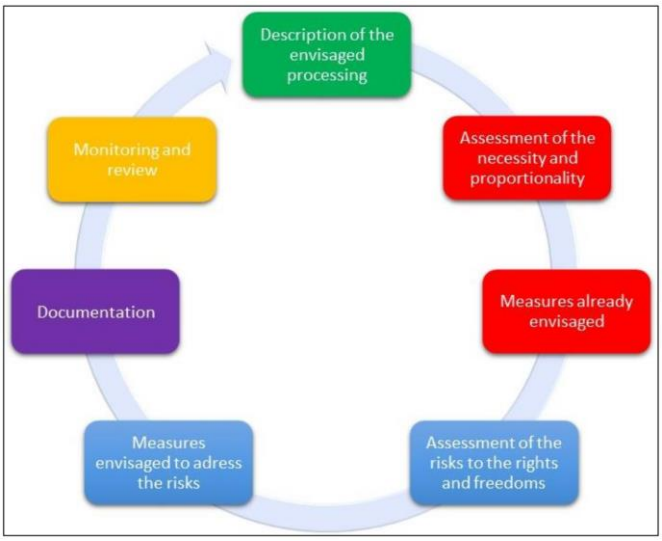
After assessing the risks, and collecting and incorporating all the inputs, the plan to mitigate the identified risks should be constructed. In cases where the identified risks cannot be sufficiently addressed by the data controller, the supervisory data protection authority must be consulted.

There are ten criteria that, according to the WP29 Opinion, should be considered when deciding whether the DPIA is necessary. Does the processing involve:

1. *Evaluation or scoring*, where examples include profiling, credit checks, building marketing profiles.
2. *Automated decisions with legal or similar effects*, where processing may lead to discrimination or exclusion of individuals.

3. *Systematic monitoring*, where processing includes monitoring of a "publicly accessible area" where the data subjects may not be aware who is and for which purpose they are collecting and processing personal data.
4. *Sensitive data of highly personal nature*, where special personal data are processed, as defined in Article 9 of the GDPR (i.e. political opinions, health information, etc.). However, for some categories of data, while still deemed sensitive, their public availability may be taken into account when assessing the risk and using these data for certain purposes.
5. *Data processed on a large scale*; the GDPR does not define what large-scale means, however WP29 mentions that factors that may be considered include the number of concerned data subjects, volume or range of items being processed, duration of processing, or the geographical extent of the processing activity.
6. *Matching or combining datasets*, where initial processing has created data on which further processing would exceed the reasonable expectation of the data subjects.
7. *Data concerning vulnerable data subjects*, where examples of such subjects are children, patients, employees, and similar.
8. *Innovative use or applying new technological or organisational solutions*, for example where processing may be combining fingerprint with facial data i.e. whenever personal or societal consequences of processing may be unknown or high.
9. *Data transfer across borders outside the European Union*, taking into consideration the country of destinations, possibility of future transfers or transfers based on derogations specified by the GDPR, among others.
10. *When processing prevents data subjects from exercising a right or using a service or a contract*, where examples may be when banks screen customers against a credit reference to decide whether to grant a loan.

The WP29 Opinion 248 (is was revised once, we consider *rev1*) states that, in most cases, when meeting two or more criteria the data controller should conduct a DPIA, regardless of the compensating measures the data controller plans to adopt. It goes further, stating that in some cases even meeting only one of these criteria may already require conducting a DPIA. And in all cases, this DPIA should then be an iterative process, and be reviewed periodically. The iterative process, as outlined in WP29 248.rev1 opinion, is shown in the graphic.



In addition, the GDPR itself sets out the minimum features of a DPIA (Article 35(7), and recitals 84 and 90). It should contain:

- “a description of the envisaged processing operations and the purposes of the processing”;
- “an assessment of the necessity and proportionality of the processing”;
- “an assessment of the risks to the rights and freedoms of data subjects”;
- “the measures envisaged to:
    - “address the risks”;
    - “protect the data”
    - “demonstrate compliance with this Regulation”.

The requirements outlined in the GDPR provide a generic, scalable framework for conducting a DPIA. And when assessing the risks one should consider them in relation to the “rights and freedoms of the natural persons”. In the recital 90, three processes are delineated:

- establishing the context: “taking into account the nature, scope, context and purposes of the processing and the sources of the risk”
- assessing the risks: “assess the particular likelihood and severity of the high risk”
- treating the risks: “mitigating that risk, ensuring the protection of personal data and demonstrating compliance with this Regulation”

As stated in the WP29.rev1 opinion: “A ‘risk’ is a scenario describing an event and its consequences, estimated in terms of severity and likelihood”.

The practical implementation, but also the initial decision whether to conduct a DPIA, will depend on the actual circumstances and requirements.

In the next chapter we will evaluate the risks that are present in the use case of research communities employing FIM. We will consider their impact, their likelihood, and which controls or mitigations can be employed to reduce or eliminate the risks, if necessary.

4. Risk assessment and DPIA impact on Community and Infrastructure Proxies

The AARC Blueprint Architecture (BPA) provides a framework for research communities to organise federated identity and access management, and structure the provisioning of access to resources and services within the community and generic e-Infrastructures. This specific framework providers sufficient commonality in the processing of personal data that a common risk assessment becomes feasible. We thus assess the risks for the Infrastructures involved within BPA use cases, evaluate the risks’ severity and likelihood, and provide generic guidance as to whether such risks are “likely to result high risk” for the data subjects. Furthermore, we will provide an overview of possible measures that, when undertaken appropriately, may lessen and mitigate the risks. Given that each research community and infrastructure is unique, no generic guidance can in itself be considered definitive (legal) advice, thus it must always be considered as input to an organisational

decision. Yet we hope that it provides the appropriate hooks and references to significantly ease such a decision.

4.1. Establishing the context

Federated Identity Management (FIM) is an “arrangement that can be made among multiple organisations that let subscribers use the same identification data to obtain access to the secured resources of all organisations in the group” [FIM4R1]. Some obvious properties of FIM have inspired its use for Infrastructures and users (convenience and simplified user management among others), but additionally, and of specific interest when assessing risks, operating within FIM principles also provides for data minimisation and better data security over conventional methods. This is achieved by technical means (specific user attributes can be requested), by policy (definition of entity categories like Research and Scholarship that promote the use of just a few ‘harmless’ attributes like organisational email and the users’ own name), and in the AARC BPA by the scoping of attributes and user information to the community (in the community attribute authorities and Infrastructure proxies) and only the involved service providers. It also has one other significant advantage: user credentials (passwords) are managed in a single trusted place (the user’s own home organisation), so that such sensitive information is not distributed throughout the services. It makes theft of such personal passwords much less of an issue - even when the user works in “dynamic collaborations that cross organisational and national boundaries”.

The development and the progress in using FIM has led to the general acceptance of proxies, “to act as a mediator between identity providers and the services used by research disciplines”. This is also recommended by the AARC Blueprint Architecture (BPA) [BPA]. The general scheme of the BPA 2016 is again shown graphically here for reference.

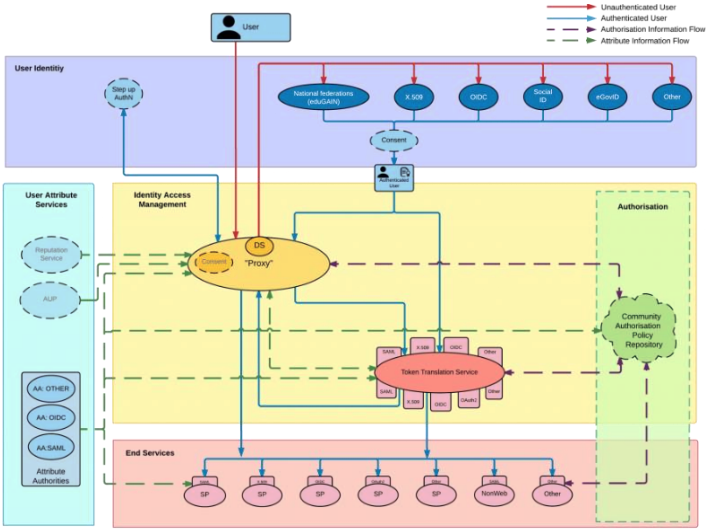


Figure 2 AARC Blueprint Architecture 2017. This is the architectural model considered for the DPIA assessment process described.

In the guidance given here, we specifically consider the BPA model and the interposition of the proxy when accessing and using services - and limit ourselves to personal data collected as a result of users accessing (using) these services, including any accounting data collected. In the initial phase, we also limit ourselves to those cases where the user of the infrastructure itself is not a sensitive issue, i.e. for research where the freedoms of the researcher in itself are not at risk. More complex cases, which could occur in biological and medical research, need additional input from the research communities involved and have to be considered at a later time.

4.2. Risk Assessment

When talking about the risks, here we are assessing the risks for the users (or data subjects) accessing and using the resources. Although these risks are different than the risks of service providers, they could be considered together for situations when risks of one is influencing the other as in, for example, penetration of services may leak data that contain personal information of the users. However, in this document we will focus on the risks for the data subjects, and consider methods for minimisation and mitigation of such risks. These are the risks that the service provider (and Infrastructure) must weigh and consider - and possibly present to the user.

Federated Identity Management and Federated Access Management (here taken together as "FIM") is recognised by the regulators as a privacy enhancing tool [CORMACK]. With FIM, the information about the users is only released when accessing the service, and then only the necessary information required by the service is released. In general, the information contained in this use case is only the users' email and name, and a non-reassigned identifier (the "Research and Scholarship" attribute set). This scenario also benefit the organisations providing the services, or Service Providers (SPs), since it provides them with assurances on users' information and allows SPs to identify and contact the user for problem resolution and user support. There are also trusted intermediaries involved in the scheme (either to just broker trust between the users' home organisation and the service provider, or to convey and augment the access with community attributes and assurance statements), which can serve both to scope the flow of personal data and to provide community-based 'pseudonymous' identifiers. This set-up reduces the information collected and processed about the users, which is in line with the data minimisation benefits previously mentioned.

The information that is processed, i.e. email and name (or identity), is considered *Common Personal Data* [CNIL-MAN]. Additional data that typically may be collected include *Connection Data* such as IP addresses, event logs, etc. These data, while still personal data, are not sensitive data such as biometric information or bank information, nor are they considered sensitive data as defined in the GDPR (Art 9).

4.3. EDBP DPIA criteria

In the BPA use cases - where data subjects' data are being processed to provide them with access to services - in order to address whether a formal process of the DPIA is necessary we should consider the ten criteria outlined by the EDPB. For the research and collaboration scenario's two of these stand out: *cross-border transfer* (beyond the EU, as research is global) and, potentially, *data processed on a large scale*. And although the EDPB as general guidance recommends that a formal DPIA be conducted when two or more criterias are met, it leaves it explicitly open to do a specific assessment and from that conclude that a DPIA will not be necessary. As per the GDPR, a DPIA is only necessary when processing is "likely to result in high risks". And such a decision can be substantiated by preliminary guidance provided by some of the more active national regulatory bodies that today provide guidelines for DPIAs.

But before delving into these guidelines we briefly consider the other criteria and their applicability to the research communities using FIM: *Evaluation or scoring* is typically not used for providing access to users, nor *systematic monitoring*, especially in the sense described by the EDPB. *Sensitive data* is also not processed, since it is not necessary to provide a service. Sensitive data are processed in some research communities, however that is outside the scope of the AARC BPA and the Infrastructure itself (communities that perform such processing will be well aware of their own need for a DPIA, e.g. for research on criminology, or personalised medicine, and like domains), nor is it the processing scenario we are considering. *Innovative use or applying technological or organisation solutions* are not necessary, and therefore not used, for the use case we're describing (this e.g. refers to machine learning and AI techniques - FIM is not 'innovative' but rather 'standard'). Also, the information collected is the minimal set to provide a service and the used technological solutions (like SAML [SAML], OIDC [OIDC], and X.509 [PKIX]) are industry standards. The remaining criteria are even less applicable.

Before assessing the severity and likelihood of the risks, and therefore providing an estimate of the risk levels (i.e. whether it will be "high" or not), we first consider the controls on proportionality and the necessity of processing - and any controls protecting data subjects' rights - and show examples on how they are addressed in our FIM and AARC BPA scenario.

Proportionality and necessity of processing

- *Purpose*: *specified, explicit, and legitimate* - In our case the purpose is to provide *access to resources*, as defined by version 2 (draft) of the GÉANT Data Protection Code of Conduct [DPCoCo]
- *Basis: lawfulness of processing, prohibition of misuse* - Under GDPR, conditions for consent are strengthened, and access to resources in the FIM environment is typically done for professional reasons. Therefore, *legitimate interest* as a legal basis is the logical choice, since, as stated by WP29, it aims for a "balanced approach, which ensures the necessary flexibility for data controllers for situations where there is no undue impact on data subjects, while at the same time providing sufficient legal certainty and guarantees to data subjects that this open-ended provision will not be misused" [WP29-217]
- *Data minimisation: adequate, relevant and limited* - As mentioned, under REFEDS "Research and Scholarship" and AARC recommendations only email and name are collected. Further information may be collected on a need basis, but additional information is typically about the assurances of the identity (i.e. how certain is that the user is who says it is, affiliation, etc.) or "freshness" of the information. This data is still considered *Common personal data*, and therefore not sensitive. The community attributes (group information, service access rights) are assigned to the user, necessary for the purpose of granting access, and not in themselves revealing information.
- *Quality of data: accurate and kept up-to-date* - In the FIM scenario, data is typically released upon each access to services. Furthermore, by policy under the GEANT DP

CoCo, service providers give the ability for the users to update or remove their information, if necessary and required to do so.

- *Storage periods: limited* - Personal data is usually removed after certain periods of time (typical interval is 6-18 months) - and Infrastructures adhering to GEANT DPCoCo as well as those part of the European e-Infrastructures, make this explicit in their policy frameworks.

Controls protecting data subjects' rights

- *Information for data subjects* - SPs (and specifically also Communities and Infrastructure Proxies on their behalf, for reasons of usability and scalability) provide a Privacy Policy to users' on their first access to services. In the policy it is explained how their data are being processed.
- *Rights to rectification and erasure* - If the users are no longer accessing services, their data is usually removed after certain periods of time. Furthermore, in the Privacy Policy is typically listed a contact for the users to address this issue.
- *Transfers* - For transferring the data outside of EU, SPs have several methods to use, and current effort for the new DP CoCo is also addressing this issue.

Types of outcomes arising from risks scenarios occurring are

- *Illegitimate access to personal data* - where outcomes could range from none, i.e. data is not used, to some actual use (regulators give examples such as spamming, etc.)
- *Unwanted modification of personal data* - where the outcome is a result of malfunction, i.e. data not used properly, and use (i.e. misuse) of data
- *Disappearance of personal data* - where the outcome is a malfunction, i.e. resulting in errors or malfunctions in using the service, and blockage, i.e. resulting in service not accessible anymore

In FIM, all three scenarios are possible, however their impact depends on their severity and likelihood, jointly an input to the estimation of the risk levels. Severity is defined as a consequence or a magnitude of risk. Its estimate is influenced by a nature of the potential impact, i.e. nature of data, data subjects, purpose of risks [CNIL-METH], etc. Likelihood express the possibility of a risk occurring.

Severity

Although in its early stages, the most extensive source of information on risk assessment, particular also considering the collaborative use cases and the context of accessing shares resources, comes from the *Commission Nationale de l'Informatique et des Libertés* (CNIL), the French Data Protection Authority. In its "PIA Knowledge Bases" white paper [CNIL-KB],

the following risks are explicitly listed as 'negligible' or 'limited', in terms of both material[1] and moral[2] impacts.

Negligible:

- Loss of time in repeating
- Spam emails
- Targeted advertising
- Mere annoyance caused by information received or requested
- Feeling of losing control of one's data
- Feeling of invasion of privacy without real or objective harm (e.g. commercial intrusion)
- Loss of time in configuring one's data

Limited:

- Unticipated payments, additional costs (e.g. bank charges)
- Denial of access to administrative or commercial services
- Lost opportunities of comfort (termination of an online account)
- Minor but objective psychological ailments (defamation)
- Feeling of invasion of privacy without irreversible damage
- Intimidation on social networks

The level of severity may be raised or lowered by including following factors:

- Level of identification of personal data
- Nature of risk sources
- Number of interconnections (especially with foreign sites)
- Number of recipients (which facilitates the correlation between originally separated personal data)

Although more severe risks most certainly do exists, they do not apply to the FIM and Infrastructure use cases that must be taken into consideration here.

So in the FIM scenario, we should consider several of the negligible or limited impact scenarios as possible and concrete risks. As previously mentioned, types of data in the considered FIM scenario are emails, names, and IP logs. When considering the severity of risks, we can identify that the risks reach at most 'limited impact' (denial of access to commercial services), and in the majority of cases are likely 'negligible' (receiving spam emails, annoyance, fear of lost or invasion of privacy without real harm). Even though the considered FIM use case is by its nature distributed and international, there is no prima facie reason to assume that merely because of this scope the severity level should significantly increase. And regardless, the FIM use case is still preferable to the non-federated alternative, since, for example, in a scenario in which a single service may experience a data breach where personal data may be compromised, due to the nature of data (i.e. emails,

[1] From CNIL: "Loss incurred or lost revenue with respect to an individual's assets"
[2] From CNIL: "Physical or emotional suffering, disfigurement or loss of amenity."

names) and the FIM's distributed nature and size (i.e. research communities are typically not very large; not all users access all services), the impact is still substantially lower than in the scenario where a large, centralized commercial service provider would experience a data breach (in which case the user base may be considerably larger, and data may include more sensitive data such as service passwords, bank or credit card information). Especially given the acknowledges fact that users tend to re-use the same credentials for many (if not all) services, not having credential (password) data distributed across the Infrastructure is a very significant advantage of FIM, making a data breach of a research service provider much less severe.

Likelihood

Likelihood represents the feasibility of a risk to occur, and the scale, taken from CNIL, is ranging from 'negligible', meaning that the considered risk source does not seem possible to materialize the threat, to 'maximum', where it is very easy to materialize the threat. Again, the level can be subsequently raised or lowered by the following factors:

- open to the Internet or it being a closed system
- data exchanges with foreign countries (or not)
- interconnections with other systems or no interconnection
- heterogeneity or homogeneity of the system
- variability or stability of the system
- the organization's image

In summary, the following table, provided by CNIL, can serve as a useful template to capture the identified threats and their level of impact. Even when not conducting the formal DPIA, this table may offer guidance for research communities in documenting their risk assessment and risk mitigation strategies, to further demonstrate compliance with the GDPR.

Risks	Impacts on data subjects	Main risk sources	Main threats	Existing or planned measures	Severity	Likelihood
Illegitimate access to personal data						
Unwanted change of data						
Disappearance of data						

4.4. Risk Mitigation

Based on the considerations above, and on the guidance from the French regulator, it is appropriate to infer that – even if data are processed on a large scale and cross national boundaries beyond the EU (something that is made explicit to the user and is actually much expected and appreciated) – the processing of personal data as a result of using the Infrastructure is unlikely to be high.

However, even though as a result there is no obvious requirement to conduct a formal DPIA in the FIM scenario research communities is, GDPR still mandates that identified risks and their mitigation should be considered. We already mentioned certain good practices in regards to the proportionality and necessity of processing, and controls protecting personal data. We will elaborate on these controls further, and provide recommendations and best practices to further lower the risks for users. This will strengthen the position of the Infrastructure and the service providers in explaining and demonstrating GDPR compliance.

Addressing risks to the rights of the data subjects is related to many different aspects of information security management, including e.g. also physical security of services. This potential issue is recognised by the research communities and consultations are ongoing in addressing them. For example, the WISE community [WISE] issued guidance on addressing security risks [WISE-RISK] and continuously revises and improves it to address newly identified risks. Furthermore, the joint community effort including both e-Infrastructures, communities, and the R&E federation operators in REFEDS [REFEDS] produced the *Security Incident Response Trust Framework for Federated Identity* (Sirtfi) [SIRTFI], which aims to enable the coordination of incident response across federated organisations. *Sirtfi* facilitates sharing of data for incident response purposes, in itself an effort specifically endorsed by the EDPB [WP29-262]. REFEDS Research and Scholarship (R&S) [REFEDS-RS] aims for the minimal release of information while still providing enough information to access services. The *Scalable Negotiator for a Community Trust Framework in Federated Infrastructures* (*Snctfi*) is an scheme to facilitate trust in the “proxied” environment promoted by the AARC BPA [SNCTFI]. Compliance ensures that services behind the proxy are following necessary security and privacy best practices. The AARC project, in cooperation with wider community and the e-Infrastructures in EOSC-Hub, is engaged in an ongoing effort to produce a comprehensive set of rules and documents (a “policy development kit”) to help communities operate their services in compliance with all the mentioned frameworks.

In estimating risks for the formal DPIA, also existing Codes of conduct should be taken into account. The GÉANT Data Protection Code of Conduct (DPCoCo) version 2 is an effort by community to ease the implementation of and expression of adherence to the requirements of the EU Data Protection Directive and of the upcoming General Data Protection Regulation (GDPR) in federated identity management, specifically also enabling the exchange of personal data outside the EU. The Data protection Code of Conduct defines behavioral rules for services that want to receive users’ data, specifies the purpose of processing (i.e. access to resources), and the measures to protect such data. It addresses the necessary controls for protecting data and proportionality and necessity of processing. Furthermore, it requests the potential “abiders” to follow best security and operational practices.

5. Summary

Providing and accessing services using Federated Identity Management (FIM) poses some risks for the rights of data subjects because of the wide scope of collaboration and its inherent global and cross-national aspects. In this report we analyse the risks in the context of the AARC Blueprint Architecture (BPA) model and the Infrastructure Proxy, and - leveraging the FIM technical mechanisms and the policy guidance for communities and Infrastructures - place these in the context of the General Data Protection Regulation and the “Data Protection Impact Assessment” identified therein as a risk assessment mechanism.

Based on the regulatory guidance available and the inherent safeguards built into the FIM model or service access, we show that significant aspects of GDPR compliance are already satisfied, specifically in data minimisation, reduction of the spread of personal data (and critical elements like credentials and passwords), and data security. Adherence to community best practice, limiting data to that based on REFEDS Research and Scholarship, and by implementation of the GEANT Data Protection Code of Conduct, Sirtfi, and the use of the Snctfi policy framework to ensure coherent behaviour of services ‘behind’ the BPA Community and Infrastructure Proxies, significantly mitigates any residual risk.

We argue that the mentioned best practices, with documented and enforced procedures, significantly reduce the risks for the data subjects. As a result, in many cases it is unlikely that a DPIA will be needed for data gathered as a result of using the Infrastructure itself. Of course, the risk assessment for the research data itself (which it outside the scope of the AAI) may well warrant such an assessment if it concerns e.g. personal medical data, criminological data, or in the special cases where the fact of doing research in itself could expose researchers to risk (e.g. for research which in parts of society or media is not so well received).

Still potential risks may still remain in specific cases. Communities and Infrastructures are can use this guidance document to inspire and guide their implementation, acknowledging that generic document such as this one can not be construed as legal advice in any particular jurisdiction.

References

AARC-G016 https://aarc-project.eu/guidelines/aarc-g016/

BPA AARC Blueprint Architecture; AARC-G012 https://aarc-project.eu/guidelines/aarc-g012/; see also https://aarc-project.eu/architecture

CNIL-MAN Commission Nationale de l'Informatique et des Libertés *DPIA manual*; https://www.cnil.fr/en/privacy-impact-assessments-cnil-publishes-its-pia-manual

CNIL-METH Commission Nationale de l'Informatique et des Libertés *CNIL PIA methodology*; https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-1-en-methodology.pdf

CNIL-KB Commission Nationale de l'Informatique et des Libertés *CNIL PIA Knowledge Base*; https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-3-en-knowledgebases.pdf

CORMACK A. Cormack *Federated Access Management and the GDPR*"; https://community.jisc.ac.uk/blogs/regulatory-developments/article/federated-access-management-and-gdpr

DPCoCo GEANT Data Protection Code of Conduct (version 2) https://wiki.refeds.org/display/CODE/Data+Protection+Code+of+Conduct+Home

FIM4R1 Federated identity management for research collaborations CERN-OPEN-2012-006, Apr 23, 2012 - https://cds.cern.ch/record/1442597

GDPR General Data Protection Regulation 2016/679 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

KUNER2018 Kuner, Christopher, *International Organizations and the EU General Data Protection Regulation* (February 1, 2018). University of Cambridge Faculty of Law Research Paper No. 20/2018. Available at SSRN: https://ssrn.com/abstract=3050675 or http://dx.doi.org/10.2139/ssrn.3050675

OIDC OpenID Connect - https://openid.net/connect/

PKIX See RFC5280 (https://tools.ietf.org/html/rfc5280); ITU X.509 - http://www.itu.int/rec/T-REC-X.509/en; and its applications in transport security and authentication

REFEDS Research and Education Federations; https://www.refeds.org/

REFEDS-RS REFEDS Research and Scholarship Entity Category; https://wiki.refeds.org/display/ENT/Research+and+Scholarship

SAML Security Assertion Markup Language (SAML) - https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security

SIRTFI Security Incident Response Framework for Federated Identity; https://refeds.org/sirtfi

SNCTFI Scalable Negotiator for a Community Trust Framework in Federated Infrastructures; https://igtf.net/snctfi/

WISE WISE Information Security for e-Infrastructures; https://wise-community.org/

WISE-RISK WISE Risk Management Template; https://wise-community.org/risk-assessment/

WP29-217 WP29 Opinion on the legitimate interest of the data controllers - http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf

WP29-248 WP29 DPIA Opinion - http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236

WP29-262 WP29 Opinion on Guidelines on Article 49 of Regulation 2016/679; http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=614232