Standardising GA4GH IAM/AAI

Goals

This working document aims to collect use cases, requirements and a critical mass of people working towards standardised Identity & Access Management (IAM) and/or Authentication & Authorization Infrastructure (AAI) policies and guidelines for federated data storage, sharing and analytics for use with GA4GH APIs in a cross-standard and potentially cross-work stream setting (~FASP).

To achieve that, we propose the following milestones:

1. **Task force:**
   Assemble task force of interested/affected parties to address subsequent milestones
2. **Requirements collection:**
   Collect use cases and IAM/AAI requirements for federated analytics from different stakeholder (the more variability, the better)
3. **Curated requirements collection:**
   Prepare curated list of IAM/AAI requirements
4. **Model:**
   Taking into account current requirements, current guidelines/specs and other proposed models (see below), do one or more (or a mixture) of the following to address the curated requirements:
   - Support a proposed model (perhaps preparing a set of questions to clarify points of doubt)
   - Extend one or more existing models/proposals
   - Develop one or more new models
5. **Position document:**
   Compile models (existing, extended or new) into a position document that describes in plain text how the models address the curated list of requirements and, if applicable, existing models or proposals fail to do so
6. **Promotion:**
   Promote position within GA4GH and other relevant communities

Resources

Existing/proposed models

- Current GA4GH specifications/guidelines on
  - Passport and Visas
  - AAI

- Proposal: Work order tokens (by NIH): https://drive.google.com/file/d/112EOHsrWpIITaK_CpEhPuoN5v0wVI_zJ/view
- Current ELIXIR::GA4GH Cloud AAI guidelines: https://github.com/elixir-cloud-aai/elixir-aai-guidelines
- User-Managed Access (UMA): https://en.wikipedia.org/wiki/User-Managed_Access
- Incremental token downscoping via Macaroons: https://en.wikipedia.org/wiki/Macaroons_(computer_science) https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/ https://www.ndss-symposium.org/ndss2014/ndss-2014-programme/macaroons-cookies-contextual-caveats-decentralized-authorization-cloud/
- Open source languages for writing AAI policies/rules: https://www.cerbos.dev/ https://openfga.dev/
- OpenID Foundation (working on similar problems): https://openid.net/wg/heart/ https://openid.net/wordpress-content/uploads/2022/06/OIDF-and-the-Health-Whitepaper-June-21.pdf

Other relevant GA4GH Specifications

- Data Repository Service (DRS) API
- Task Execution Service (TES) API
- Workflow Executions Service (WES) API
- Service Registry API
- Tool Registry Service (TRS) API
- Crypt4GH

Relevant GA4GH Meetings and Notes

- Cloud WS (every other Mon)
- FASP (1st & 3rd Mon of the month, different time zones)
- Passports/AAI Technical working Subgroup (every Thu)
- Passports workstream (Wed, every 6 weeks)
- ELIXIR::GA4GH AAI Technical Call agenda & minutes (monthly on 2nd Mon)

Contributing

There are multiple ways in which you can contribute to this effort:

- Join the Task Force
- Write down the way in which you or your organisation envisions to use federated analytics by specifying a Use Case

- Write down individual AIM/AAI Requirements that are important to you or your organisation; requirements are ideally, but necessarily, derived from a use case (if not, just put N/A in the “Use case” field)

Task force members

Please consider signing up for the task force. The more we are, the easier it will become to find a solution and ensure widespread adoption. Just sign up in the list below and try to attend the monthly meetings (see Goals section).

Name	Email	Affiliation	ORCID	GitHub
Alexander Kanitz	alexanderkanitz@gmail.com	University of Basel, Switzerland; Swiss Institute of Bioinformatics; ELIXIR-CH / Pacific Analytics Pty Ltd	https://orcid.org/0000-0002-3468-0652	https://github.com/uniqueg
Martin Kuba	makub@ics.muni.cz	Masaryk University, Brno, Czechia; CESNET	https://orcid.org/0000-0002-0305-7446	https://github.com/martin-kuba
Geoff Coles	geoff.coles@genomicsengland.co.uk	Genomics England UK	N/A	https://github.com/geoffcoles
Philip R. Kensche	p.kensche@dkfz.de	German Cancer Research Center (DKFZ), Heidelberg, Germany	https://orcid.org/0000-0003-1299-9600	https://github.com/vinjana
Johan Viklund	johan.viklund@nbis.se	National Bioinformatic Infrastructure Sweden, Uppsala University, Uppsala, Sweden	https://orcid.org/0000-0003-1984-8522	https://github.com/viklund
Dylan Spalding	dylan.spalding@csc.fi	CSC, Finland	https://orcid.org/my-orcid?orcid=0000-0002-4285-2493	https://github.com/jdylan

Venkat Malladi	vmalladi@microst.com	Microsoft	https://orcid.org/0000-0002-0144-0564	https://github.com/vsmalladi
Takudzwa Musarurwa	tnmusarurwa@gmail.com	eLwazi ODSP, University of Cape Town	https://orcid.org/0000-0002-0138-0602	https://github.com/takudzwa-coder
Pavel Nikonorov	pavel@genxt.network	GENXT, Hinxton, UK	https://orcid.org/my-orcid?orcid=0000-0002-8471-2069	https://github.com/pavelnikonorov
Karen Cranston	karen.cranston@gmail.com	Pan Canadian Genome Library, University Health Network, Canada	https://orcid.org/0000-0002-4798-9499	https://github.com/kcranston
Heidi Sofia	Heidi.Sofia@nih.gov	National Center for Biotechnology Information, NIH		

Use cases

Describe your use case in some detail. Add more rows if necessary. Please include the organisation you are representing and one or more contacts.

Identifier	Description	Organisation	Contact

UC01	A user triggers a workflow run that is processed by a GA4GH WES service, which in turn sends out task execution requests across a network of GA4GH TES services; both WES and TES services may need to access publicly available data from various storage solutions (s3, HTTPS, FTP), as well as data objects behind GA4GH DRS services, some or all of which may be encrypted (e.g., by Crypt4GH). Which actual WES and TES (and possibly even DRS) services will be used may be dynamically decided after querying a GA4GH Service Registry for available candidates, considering data transfer/usage restrictions. In the case of DRS inputs, only data can be used to which a user has been granted access by the relevant authorities. For any other data, as well as for compute backends and cryptographic tools, credentials/tokens/keys must be supplied along with the workflow run request, as required by the service used. Additional service layers (e.g., gateways, or service types not mentioned or not yet defined) may be used along the way, if required for fulfilling the user's request and if adhering to security guidelines, and all services may be operated by different service providers that may operate from within different jurisdictions. Upon the completion of requests, any output data must be made available to the user in a secure, encrypted way. Additional security measures may apply, as requested by policies associated with individual resources (data or other), e.g., that computations can only be computed in Trusted Execution Environments, offline and/or encryption at rest, in transit and/or in use policies, limitations to which tools/workflows can be used for particular data sets or compute environments etc.	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
UC02	The client side formed by the four APIs of Cloud WorkStream, WES, TES, TRS and DRS, respectively constitute the four components of Bio-OS's control layer, execution layer, Tool layer and data layer. The four components are deployed in different sites in a distributed form, and each site can deploy all four or at least one of them to join the Bio-OS Network according to the actual situation. In the secondary analysis stage, the user can access any desired control panel from the Bio-OS Network, specify the WDL from any tool layer server with TRS ID, and push analysis task to any execution layer server. The execution layer server pulls the docker image from any tool layer server using TRS ID, and pulls the data from any data layer server using DRS ID for analysis and returns the result. The pulling process of WDL, docker image and data requires permission authentication. The	GCBI & GZNL（China）	liu_jilong@gzlab.ac.cn

	response of the execution layer server requires permission authentication for the task. In the tertiary analysis stage, the user can access from any desired control layer server on the Bio-OS Network, and pull the specified docker image from any specified tool layer server on the Bio-OS Network to start Jupyter Notebook execute at any specified execution layer server. The Jupyter server can pull data from the data layer server with DRS ID, and the calculation result is returned to the control layer. The whole process needs to be authenticated the same as the secondary analysis.		
UC03	Using ga4GH APIs as a standard for all tools and researchers to both discover and access genomics england data and to run workflows Enabling data federation and federated workflow execution through the use of GA4GH apis	Genomics England	geoff.coles@genomicsengland.co.uk
UC04	Authentication data should not by design (lack of protocol expressiveness) have to be written to input files or logs. For instance, it should not be necessary to write tokens for accessing DRSs, TRSs, and TESs, which are used by a workflow, into the workflow's configuration file. Instead it should be possible to keep authentication information in memory, and provide the tokens for multiple DRSs/TRSs/TESs/etc. via the WES, and only channel this information into the engine process (e.g. at startup). (Note: It might be acceptable to store authentication data into an encrypted keystore and forward only the keystore password securely/in memory. It really depends on our threat-model.)	WESkit in cloud deployment	p.kensche@dkfz.de
UC05	Assume no single sign-on (SSO) is possible, e.g. because of a cross-organizational project for which no SSO could be organised. Still a workflow should analyse data at multiple sites and compile the results. To collect the data it is necessary to define a token for each organisational unit's DRS or S3 storage. This use-case can be extended, if the workflow not just collects data (e.g. to process it locally), but if the workflow does have to analyse the data remotely, e.g. to aggregate data remotely to protect patient rights. For this it is necessary to provide tokens to e.g. TES and maybe TRS services at each site. This additional requirement can be avoided, if we assume that at least each site has an established local SSO, i.e. all relevant services can be accessed with the same token.	WESkit in cloud deployment	p.kensche@dkfz.de

UC06	**System used for processing data** (TES) would have a **specification** (application, capacity, geographical location and also data security..) **and identification** to control that **certain data** can be exposed only for the selected system (identifier) or systems compliant some features. Thus, some kind of Visa_System. Visa_User would have an attribute, which requires that the Visa_System is also required? User would also indicate the analysis type (intended data use). In Bigpicture it might be set in REMS? How dataset requirement is specified? How DRS requires specification for target system? Client – (Visa_System,Visa_User) -> Repository Client <- (Dataset access) – Repository One related/sub use case is when a dataset is granted only for **indirect** access only. The granter would know that the system does not give direct access for the user. This might be an attribute. The software run on the system (application) would be verified. - User -> REMS->Permissions for User (+System X does user need to specify and if yes what?) for a indirect dataset X - User -> System X -> request data from repository with Visas (user Y+ system X)	Trusted environment cases. CSC involved in different projects.	jarno.laitinen@csc.fi dylan.spalding@csc.fi
UC07			
UC08			
UC09			
UC10			
UC11			
UC12			
UC13			
UC14			
UC15			
UC16			

UC17			
UC18			
UC19			
UC20			

Individual IAM/AAI requirements

Describe your requirements. Add more rows if necessary. Please include the organisation you are representing and one or more contacts.

Identifier	Use case	Description	Organisation	Contact
RQ01	UC01	GDPR compliance; among other things, natural persons (typically users) and legal persons (organisations, service providers) need to be able to 1. fully remove any digital artefacts that were created on their behalf from databases and harddrives of all services and the corresponding service providers 2. authorise service providers and individual services to perform actions on their behalf 3. revoke authorisations to service providers and services to perform actions on their behalf, taking immediate effect 4. (inasmuch as not already covered by 1.-3.) limit the list of service providers and services that they accept requests from or send requests to	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ02	UC01	Once started, actions must generally be able to complete without further user interaction; exceptions must be rare and justifiable (e.g,. highly unfavourable risk/benefit ratio) so as not to severely impact adoption. For example, users must be able to trigger workflows runs that may last for days, weeks or even months, and the corresponding workflow engines must be able to submit tasks to third party compute backends, which may in turn need to request access to data, during the entire runtime, all without requiring the user to reauthorize such requests	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ03	UC01	The list of services actually used for a given user action may not be known at the time when a user triggers that action, i.e., suitable services may be discovered and picked at runtime, provided that the chosen services are authorised by the user, as well as all upstream services, to perform the required action	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com

RQ04	UC01	Any given service should only be able/allowed to perform the minimal action required by the service to fulfil the user's request as per the request's associated authorisations; in particular, attempts should be made to limit the possibility of replay attacks and generally reduce the risk introduced by malicious actors infiltrating the network, e.g., through narrowly scoped single- or limited-use tokens	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ05	UC01	In case of failures, services should be able to retry actions as per the user's preferences	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ06	UC01	A detailed audit trail should be stored for each user request, listing all services calls associated with that request	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ07	UC01	Service providers and resources may require specific security measures to be in place in order to fulfil specific requests (e.g., 2FA/MFA or specific identification measures, such as biometrics; availability of Trusted Execution Environments on compute backends)	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ08	UC01	Authorization tokens should be signed, encrypted and be small enough to fit in headers and follow OAuth protocol	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ09	UC01	Routes should be available for users or services to dynamically provide credentials, public keys etc. that allow (other) services to, e.g., access storage solutions for uploading intermediate or final results, decrypt data (e.g., Crypt4GH)	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ10	UC01	Authorization should be able to be evaluated in an offline scenario for computing on highly sensitive data	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ11	UC01	The use of gateways (e.g., a WES that relays to another WES) should not be restricted by the security model, as long as all used services are authorised by the user and all upstream services.	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ12	UC01	Services (WES or TES) and DRS data resources may restrict the workflows and tools with which data can be processed (e.g., only individual workflows or containers, or all workflow/containers certified by some authority)	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ12	UC01	AAI guidelines and API schemas must be standardised in such a way that individual cloud/service providers can fully operate.	ELIXIR Cloud & AAI	alexanderkanitz@gmail.com
RQ13	UC02	The pulling process of WDL, docker image and data requires permission authentication.	GCBI & GZNL（	liu_jilong@gzlab.ac.cn

			China）	
RQ14	UC02	Authentication/authorization failure caused by network problems during distributed deployment Support retry	GCBI & GZNL（ China）	liu_jilong@ gzlab.ac.cn
RQ15	UC02	The authority to access Network data through DRS should distinguish between different granularities, so that the main account can assign data access permissions of different granularities to sub-accounts.	GCBI & GZNL（ China）	liu_jilong@ gzlab.ac.cn
RQ16	UC02	After the login authentication of a node control layer in the Bio-OS network, it can be authorized to push computing tasks to other Bio-OS network nodes	GCBI & GZNL（ China）	liu_jilong@ gzlab.ac.cn
RQ17	UC02	After one of the Bio-OS execution layers completes the computing task, the computing result can be authorized to be accessed/copied by other node of Bio-OS network	GCBI & GZNL（ China）	liu_jilong@ gzlab.ac.cn
RQ18	UC02	Using GA4GH WS standard services through Jupyter notebook to access multiple data repositories and support federated analysis scenarios	GCBI & GZNL（ China）	liu_jilong@ gzlab.ac.cn
RQ19				
RQ20				
RQ21				
RQ22				
RQ23				
RQ24				
RQ25				
RQ26				
RQ27				
RQ28				
RQ29				
RQ30				
RQ31				
RQ32				
RQ33				
RQ34				

RQ35				
RQ36				
RQ37				
RQ38				
RQ39				
RQ40				

Curated list of IAM/AAI requirements

TASK FORCE: To be extracted from use cases and requirements supplied by stakeholders.

Identifier	Description	Organisation	Contact
CRQ01			
CRQ02			
CRQ03			
CRQ04			
CRQ05			
CRQ06			
CRQ07			
CRQ08			
CRQ09			
CRQ10			

Models

TASK FORCE: To be developed by the task force based on the curated list of IAM/AAI requirements, taking into account existing models and proposals. More than one model may be developed. New models do not have to be prepared if existing models/proposals are sufficient to cover all requirements.

Model 1: Work order tokens

Model 2

Position

TASK FORCE: To be developed by the task force based on the suggested models. Describes in plain text how the models proposed here address the curated requirements, as well as any gaps. Explains why new models/extensions (if proposed) are necessary by addressing shortcomings of existing/proposed models. Provides security assessment of different proposed models. Taking these information together, makes a final decision on which model is preferred. Could/should be made into official GA4GH policies (or extend existing ones) and/or a white paper.