ECAS/ENES

Short description	Climate Science
Type of community	Thematic Services
Community contact	Tobias Weigel
Interviewer	Shaun de Witt
Date of interview	2018-04-19
Meetings	
Supporters	

User stories

Instruction

Requirements are based on a user story, which is is an informal, natural language description of one or more features of a software system. User stories are often written from the perspective of an end user or user of a system. Depending on the community, user stories may be written by various stakeholders including clients, users, managers or development team members. They facilitate sensemaking and communication, that is, they help software teams organize their understanding of the system and its context. Please do not confuse user story with system requirements. A user story is an informal description of a feature; a requirement is a formal description of need (See section later).

User stories may follow one of several formats or templates. The most common would be:

"As a <role>, I want <capability> so that <receive benefit>"

"In order to <receive benefit> as a <role>, I want <goal/desire>"

"As <persona>, I want <what?> so that <why?>" where a persona is a fictional stakeholder (e.g. user). A persona may include a name, picture; characteristics, behaviours, attitudes, and a goal which the product should help them achieve.

Example:

“As provider of the Climate gateway I want to empower researchers from academia to interact with datasets stored in the Climate Catalogue, and bring their own applications to analyse this data on remote cloud servers offered via EGI.”

No.	User stories
US1	As a user I want to be able to perform detailed analysis on large volumes of data in parallel using scalable cloud resource in order to achieve more rapid results than sequential processing and avoiding downloading large quantities of data to local storage.
US2	As a user I want the results of my analysis available to me anywhere and be able to share it with colleagues before publishing in order to discuss and confirm the outcomes.
US3	As a user I want to ensure my input data is accessible regardless of physical location (for example, by making use of persistent identifiers), since then I do not need to implement my own code to deal with these changes.
US4	As provider of the Climate gateway I want to empower researchers from academia to interact with datasets stored in the Climate Catalogue, and bring their own applications to analyse this data on remote cloud servers
US5	As a data producer I would like scientists to be able to reference the source data used for downstream analysis and get accreditation in any subsequent publications.
US6	As a data manager I want any analysis to generate provenance metadata in order to understand what analysis has been performed to allow both confirmation of results and increase confidence in the scientific methods and analysis.
US7	As a decision maker I want to have confidence in the scientific results on which I rely to make policy decisions.
US8	As a research infrastructure provider I would like to (*link up the community AAI portal/ensure my users only need to use a single EOSC portal to interact with ECAS*) so that my users can still use the portal they are familiar with to access resources outside of the community

US9	As a user, I want to access ECAS from my familiar community portal or the workflow I am used to without having to make use of additional services that I first would have to learn about in order to make use of ECAS.
~~**US10**~~	~~As a data manager I want to ensure data is replicated across multiple sites to ensure that it is always available to users and to ensure the safe keeping of the data.~~ (This is not relevant to ECAS - ECAS is not doing data replication ...?)
~~**US11**~~	~~As a data manager I want to ensure that data stored in external storage is safe by ensuring and auditing of regular fixity checking and, where necesary, self healing from good replicas ensures that data accessed by users is valid..~~ (Same as above, not an ECAS concern?)
US12	As an infrastructure manager I want to reduce the effort of maintaining client side code support
US13	A user would like to run a climate data analysis experiment across CMIP51 or CMIP62 data. The targeted model output (*?input?*) data come from multiple modelling groups across the globe and are therefore hosted at different ENES data sites across Europe. For a specific target experiment, as a preliminary step, the user runs a distributed search on the ENES data nodes to discover the required input files, which will result in a list of input dataset PIDs. The user then assembles a processing job specification and submits it to the ENES data analytics service. The service will arrange for the data to be available at the processing site; data locality will be exploited by default, but data movement could be also needed. Once the data are available, a data analytics workflow runs on the service instance. After the analysis has completed, the user receives an e-mail notification with a link to a shared folder at a publication service folder from which she can access and download the processing results. The challenge is to do the server-side and parallel computation on the distributed data by making transparent the access to the data (including data movement), the allocation and use of the computational resources, the analytics experiment, the provenance tracking, and the overall experiment orchestration.
US14	At a later point, another user discovers the data from the previous analysis on the publication service. In order to be certain that the data can be used for the purpose at hand, she would like to evaluate its generation history, including the details of the processing that was done and the specifics of the input data used. She retrieves data and accompanying metadata from the publication service and uses the PIDs of the processed data to discover additional information on the processing and the full list of PIDs of all input datasets. She can then make a judgement based on the assembled information whether the data are fit for her purpose without having to contact the user who originally requested the data processing.
US15	As a user, I want to be able to make selected results of my data analysis or the analysis script I developed available to others. These recipients may be my immediate colleagues but also a wider range of external third parties. The workflow to make these data available should be largely hassle-free for me.

Note: US12 and 13 come from the original ECAS proposal, the others were derived from information on the ECAS confluence page and discussions with Tobias Weigel

Use cases

Instruction

A use case is a list of actions or event steps typically defining the interactions between a role (known in the Unified Modeling Language as an actor) and a system to achieve a goal.

Include in this section any diagrams that could facilitate the understanding of the use cases and their relationships.

Step	Description of action	Dependency on 3rd party services (EOSC-hub or other)
UC1	User needs to discover the location of all required input data	ESGF Metadata Service /B2FIND
UC2	Input data must have a PID associated with it.	Community solutions assigning PIDs, possibly via B2HANDLE
UC3	ENES Data Analytics Service must be able to transfer data from its current location to the processing site based on PID (Low priority - I am not sure if we will do this; it is not entirely in the original plan, though I agree it makes sense. It depends on how data input integration ultimately looks like and what can be done with limited effort.)	gridFTP/other?
UC4	Output data must be moved to a site where users can share it for others so they can access it via a link provided by the ECAS system.	B2DROP

UC5	Users will need to register to use the ECAS service	Appropriate EOSC-AAI Solution
UC6	Data must be movable between the output storage in UC4 to a data publication service, where it must be given appropriate metadata and a PID	B2SHARE
UC7	Output data shall have appropriate and sufficient metadata and provenance information associated to enable other users to have trust in the data.	ECAS, B2HANDLE profiles (possibly their usage by B2DROP)
UC8	A link between the output data and the sources must be maintained, in addition to provenance information related to the processing steps.	ECAS, B2HANDLE profiles (possibly their usage by B2DROP)
UC9	Input data must be accessible to the computation regardless of location.	B2HANDLE usage by communities and the DataHub. Support for B2HANDLE PID profiles by DataHub.
UC10	Published output data must be assigned a PID	B2SHARE, DataHub
UC11	The provenance information must be accessible for published output data	B2SHARE & DataHub usage of B2HANDLE profiles
UC12	Users will select individual files or entire directories from their ECAS workspace and then select to publish them. The ECAS workspace will inquire a destination location for the files in the user's B2DROP workspace. The publishing workflow for the users will start from the ECAS workspace but end with a view on the publishing repository (B2DROP) showing the newly published files as confirmation.	B2DROP

Requirements

Technical Requirements

Requirement ID	EOSC-hub service	GAP (Yes/No) + description	Requirement description	Source Use Case	Related tickets
Example	EOSC-hub AAI	Yes: EOSC-hub AAI doesn't support the Marine IdP	EOSC-hub AAI should accept Marine IDs	UC1	
RQ1	EOSC-hub AAI	ESGF AAI not integrated to any AAI services	Integration of ESGF AAI to one of EOSC AAI services	UC5	EOSC WP 10-41 - Jira .

RQ2	B2DROP	Can be a central service; no need for local installation. User has no interface to B2DROP filesystem; currently user log in to jupyter with username and password. Files automatically moved to B2DROP without user intervention. GAP: Need to integrate AAI to B2DROP . For training purposes, consider using a proxy user for training purposes.	Need to be able to write directly to B2DROP (via mount point inaccessible to users), or have the workflow copy data in using NextCloud OpenCloudMesh API. Will require separate instances for training and production	UC4	
RQ2.1	B2DROP	Publishing files from an ECAS workspace to B2DROP will not require the user to log in to B2DROP separately. Aside from selecting files to publish and a destination folder, the user should also not be asked for additional information (e.g. metadata).	B2DROP must be able to understand and accept IAM security tokens provided by ECAS. Possibly additional detail questions to clarify wrt session management (transparent authentication, selecting destination folder, initiating and confirming transfer as one seamless workflow).	UC12	
RQ3	B2DROP	GAP - UNSURE - If data is moved using OpenCloudMesh, the security needs to be considered. NextCloud website recommends using SSL since user information is passed in plain text. Need to check how B2DROP is configured.	B2DROP must run with SSL enabled	UC4	
RQ4	B2SHARE	GAP - NO (if RQ2 is satified), YES (otherwise) Enable users to push files to B2SHARE. If RQ2 works there is no gap to deal with as the bridge exists. Unless RQ2 works, then beed to integrate AAI to B2SHARE	B2DROP/B2SHARE Bridge required	UC6	
RQ5	Datahub	GAP - UNCLEAR Data publishing and data ingest. Allows contacting multiple communities.			EO SC WP 10-67 - Jira . EO SC WP 10-45 - Jira .

RQ6	B2HANDLE	GAP - UNCLEAR Both input data and published derived data must be assigned a PID. For third-party users to access provenance information, B2SHARE and possibly also B2DROP need to support recording of minimal provenance information, possibly organized via B2HANDLE profiles.		UC7, UC8, UC11	

Capacity Requirements

EOSC-hub services	Amount of requested resources	Time period	Related tickets
B2DROP	Testing & Training (200GB, <1GB/file)	M5 onwards	
B2DROP	Production (see table, >500MB)	M7 onwards	
B2SHARE	Testing & Training ((200GB, <1GB/file)	M5 onwards	
B2SHARE	Production (see table, >500MB)	M7 ideally, M12 latest	
B2HANDLE	Production, 2-4 prefixes required (CMCC, DKRZ, EGI, spare)	M15 -	
DataHub	Unknow	M15	
IM/Orchestrator	Unknown	M18-	EOSCWP10-68 - Jira .