User activity white paper

Contents

About retrospective collection of user activity data

The project was initiated at the start of year 2013 inspired by idea of services design and by the results of the ENGAGE consortium members' survey on data sharing experiences. The goal was to find out if user activity data could be *illustrative* for the nature of the collaboration (Ferry, 2013).

We were also hoping that it is the lack and incompleteness of such data that will draw attention to the importance of gathering of user activity data from the very start of a collaboration. If collected from the beginning, such datasets can be used for the purposes of designing the ICT (Information and

Communications Technology) services for scientific consortia or for the purposes of development of good collaborative practices (Ferry, 2013).

We have collected and studied the user activity information from two scientific consortia: ENGAGE (www.euengage.org) and EGG (egg-consortium.org) Data harvest was possible because the same information management system, SIMBioMS (simbioms.org), was used to collect and store genetic research data for both consortia. That data storage system was first applied to serve the needs of the ENGAGE consortium and later on it was found to be also useful for the EGG consortium.

Definition of user activity analysis in ENGAGE and EGG

First, we looked at all possible data sources that would tell us about what the users actually did during the collaboration. We checked the servers and database logs in order to be sure that there is enough data to study. One of our collaborators, Massimo Menichinelli (openp2pdesign.org) proposed to build a model, find a correlation between the data sets and try to find the best visualization method, whilst originally initially we planned to visualize the user data after finding out what we want to illustrate. Other suggestions included building a presentation of the consortia ecosystem and figuring out roles and level of participation.

In addition, we wanted to show how complex the ecosystem and the interactions between users are and that they are subject to constant change. One possible way of doing so was to look at the processes and procedures followed by various actors: scientific administrator, researchers, software engineers, and at their evolution during the consortium succession.

For the sake of simplicity we chose the ENGAGE data submission workflow as a representation of initial services design (Supplementary Figure 1: Engage Data Submission process.). That process of data submission was to be compared to how the actual data upload happened. The purpose for the comparison was to:

1) Measure how successful the original design was

2) Show that by using user feedback, measuring user activity and re-modeling the ecosystem, the end result can be better fitted to the needs of all collaborators

3) Use the changing set of requirements for the ENGAGE IT infrastructure as an illustration of the iterative design cycle

4) Suggest a novel *approach* for construction of data management services collaboratively: with all the key actors involved in the design of *both* the software and the services.

Data sources

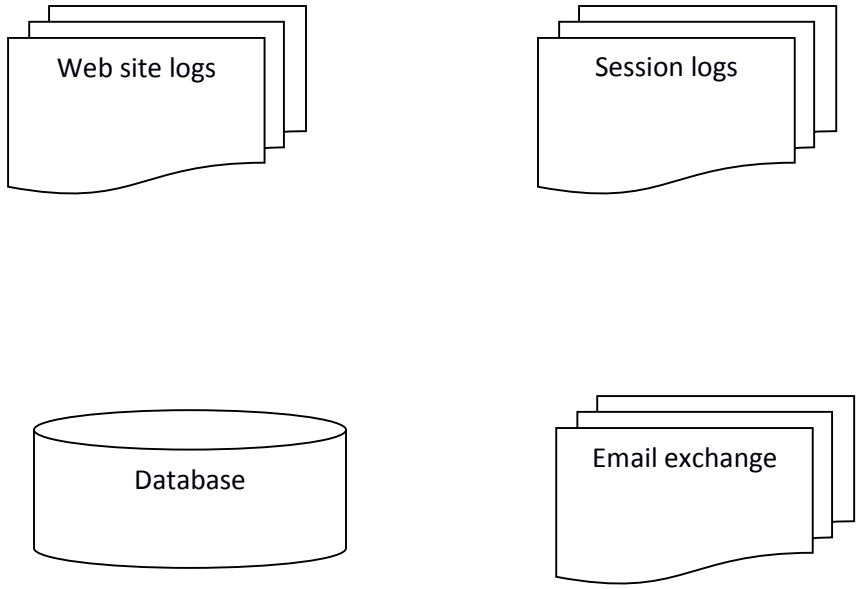


Figure 1: Data sources

1) Emails (content and statistics) that were exchanged between the consortium scientific administrator and the Primary Investigators (With permission from administration)

2) Web site log files (amount of traffic to different sites on submission periods) and the ENGAGE members area logs (a lot of information about timing of phases of the consortium)

3) Software databases (Creation/Access dates for studies, groups and users), user session log files for Assay Information System module of SIMBioMS (AIMS database). We discovered those when looking for data sources. After contacting the developer of this software it turned out that it was possible to measure real user activity time by parsing the log files.

Methods of data extraction

Once the data sources had been selected, the data was to be converted into an understandable format. The data was extracted from AIMS database (see detailed description below) with various shell scripts written in Python and Bash.

Assay Information System

Database

The first data source that was found to be suitable was the PostgreSQL database storing the relational database tables of user data. Information that we found relevant for our research was: user, group, and project creation time. Other data that was found to be useful were the pointers to actual data files in the server storage system. Data was extracted using SQL queries and exporting the results to text or csv files. Further processing included creating scripts to calculate how many users/studies/meta-projects were created for each consortium for half year periods.

Web Interface

Some of the data was mined through the AIMS web interface itself by copying and pasting to spreadsheet. This data could also have been acquired from the relational database.

Log files

Log files in themselves could not be used as they were. We needed to develop methods to extract meaningful data out of them. One of these methods was for creating a script which calculated session durations. The script worked the following way: files that contained information compiled from the session log files in a format where every session had its starting time and ending time was read to a stack data structure and converted to Python `datetime` format. Then the difference between items was calculated by removing items from stack, starting from the ending time of the last session and ending to the start time of the first session. That meant that since the prepared file with start and end times had an even number of items, session durations could be calculated with a simple algorithm:

1. Take item in the top of stack and convert it to datetime and store it to a variable t1
2. Take next item and convert it to datetime and store it to another variable t2
3. Calculate difference t2 –t1
4. Write result to file
5. Proceed until stack is empty

File system

The actual data files which AIMS database points to are stored within the server file system. After finding out how the names of these files are constructed with knowledge acquired from the developers of AIMS software, we could calculate the creation dates of the files and the volumes of data uploaded

per session and per time period. To do the mapping we wrote some SQL queries to extract file names from the database. The information was then used in a shell script that did the calculation of actual file volumes.

Activity mapping

We attempted to find out how active each AIMS institutional user was during data submission and how active they were during the data analysis. We assumed that if a user is a data submitter, he/she will optimally try to store as high volume of files in as short time as possible. On the other hand if user is a data analyst, they would download data as much as possible in as short time as possible. Having data about activity durations and file volumes was simply not enough. We had to map the user accounts to the real users of the system. AIMS did not have information on the user's true identity in a certifiable way stored inside of the system. So the user mapping had to be done by inquiring the ENGAGE scientific administrator from whom we received a list of users which were planned to be "institutional users" for the accounts. Unfortunately, there was not much that could be done to be certain that these users were actually the persons doing data submission. It was helpful that AIMS database had categories of records called "collection" or "technology". Under these categories genomic meta-studies were registered. Every collection (or meta-study) consists of different sets of summary level data on which meta-analysis is run. Every consortium partner, except the one who does the analysis, would submit their summary results under this collection.

So we acquired lists of publications which were produced with data that was exchanged through AIMS. For this to happen, we contacted the ENGAGE scientific administrator who knew which papers were produced with AIMS and which were produced by other means of sharing, like secure FTP transfers. In the case of EGG, we got the listing of papers which were produced by AIMS from one of the EGG Principal Investigators.

Once we knew which papers were produced using AIMS, we could find from the collection (meta-study) name and data files that were contained there which users had been providing data. And the mapping between AIMS institutional user accounts and the users who were accessing AIMS with these accounts, allowed to link the data to the lists of authors in consortium manuscripts. We also took into account that the position in the author list was indicative of the possible role of the user during data submission: junior authors or technical leads were often in the beginning of the list and senior researchers were at the end of the list. And taking into account the position had further refined the number of the users who could be possible data submitters.

After the list of users was finalized, we looked at how much each AIMS user had contributed to the data submission by calculating submitted data per user and per meta-project. At the same time, the meta-studies were mapped to papers (using the information we had gotten from PIs and scientific administrators), and so it as possible to measure roughly how much each consortium partner provided and how active they had been in respect to data they had provided. For further confirmation of the phasing of the project(s) we obtained a permission from the scientific administration to access email

notifications from the phase when the meta-analysis was initially proposed and later when it was initialized. That information enabled us to understand what kind of activity in the system for the given user could have been, for example if there was an indication of the start of meta-analysis in email exchange and right after that a new project was established in data management system and user activity log would show data uploads, it would strongly indicate that institutional user had been uploading data for sharing.

AIMS had a mechanism for attaching users to groups and defining access to data for the group. So in general, there was an approach for establishing a group for each meta-project (called “submitters”), members of which would only be able to upload the data and another group who could access every users data under the meta-project (called analysts).

So if a user who was an analyst was active in period of time when we see from emails that meta-analysis was proposed and proposed analysis was accepted by consortium steering board, we can deduce that high activity during that period of time means that the analyst was setting things up in AIMS. Later on, if we see that data submission has ended and the analyst account has long session durations is likely that they are downloading summary level data which submitters have provided during the data submission. For submitters we can say that their activity should peak after announcement of data submission. In their case it is desirable to have as short sessions as possible and as high data throughput as possible. Indication of problems can be that the user does not know how to use the system, which would show up in the ticketing system which was set up during the consortium, on emails to a scientific administrator which were then forwarded to system administrators, or on infrastructure behind the data submission (physical problems in connection, hardware or software failures).

The mapping enabled us to start to examine the amount of activity in data analysis and submission related to the position in the final paper. To be precise we attempted to measure the proportional credit that a researcher got from the work he or she did. We are aware that this way of measuring is not ideal and does not show activity that happened outside of the data management system. But to start measuring would be a step towards having fair way of giving credit to researchers who do the actual work, if the work flow could be made more digitally transparent. If the workspace and daily collaboration happens in environment where actions can be tracked, good scientific collaborative practices can be preserved and suggested to future consortia.

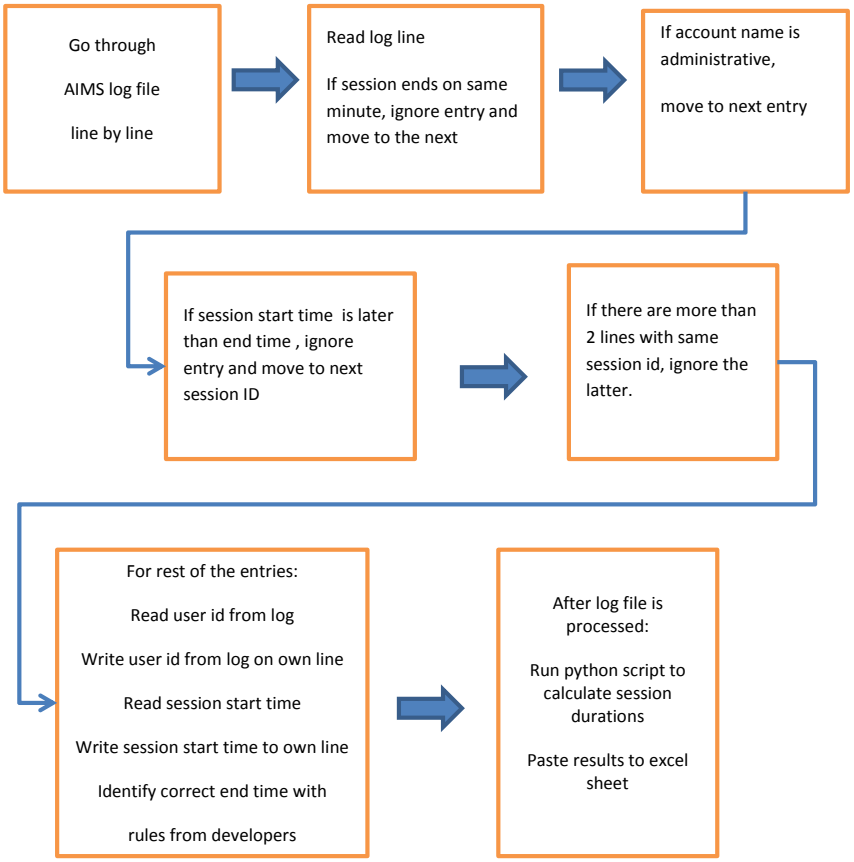


Figure 2: Log file parsing.

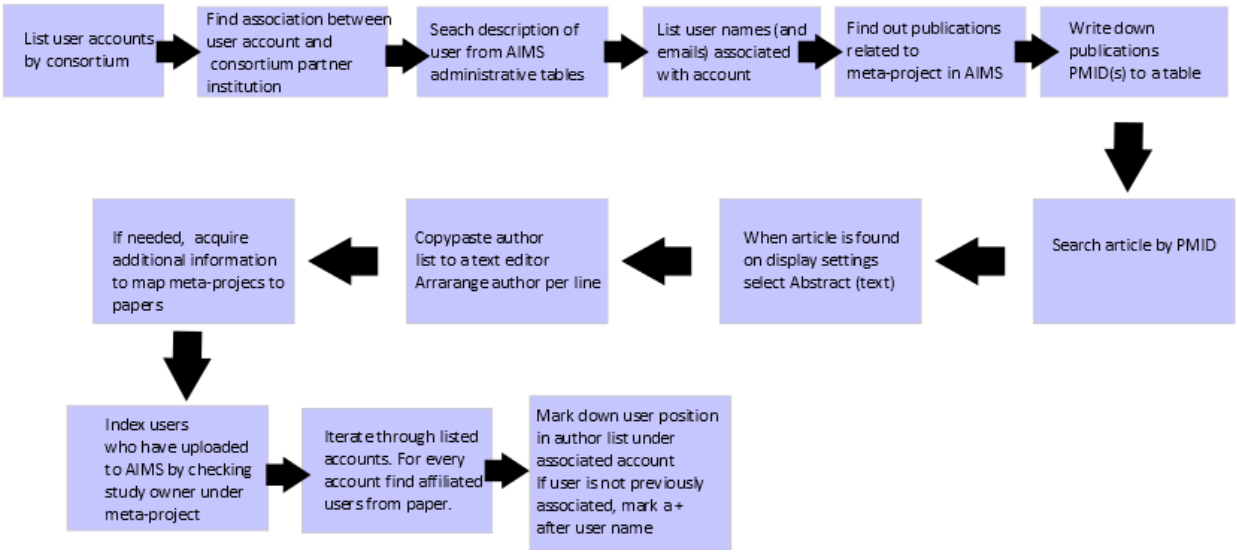


Figure 3: User – Paper mapping process.

Methodology

Our starting point for studying how services evolved in the ENGAGE consortium was the ENGAGE data submission diagram. We presumed that there had not been a systematic design of services during the consortium lifespan. So we would have a point of reference to future consortia when service design would be utilized.

During the search and selection phase we contacted experts in collaborative design as well as specialists in various knowledge domains (data visualization, human genomics researchers, designers of collaborative scientific communication, scientific administrators, PI's involved in the consortium data management design, and persons who were dealing with ethical and legal issues when setting up the ENGAGE consortium). These individuals provided us with needed information about the initiation phase of ENGAGE as well as how the information systems were developed. Likewise designers gave insight on how to start with the procedure itself, how to figure out what kind of data is useful, and did we have enough of it to start the work.

Background information

ENGAGE (European Network for Genetic and Genomic Epidemiology) was a research project funded with 12 million euros by the European Commission under the 7th Framework Programme-Health Theme with 24 partners from 13 countries with data from over 80,000 genome-wide association scans and DNA and serum/plasma samples from over 600,000 individuals. The project duration was five years, starting from January 1st, 2008 and ending in January 2013.

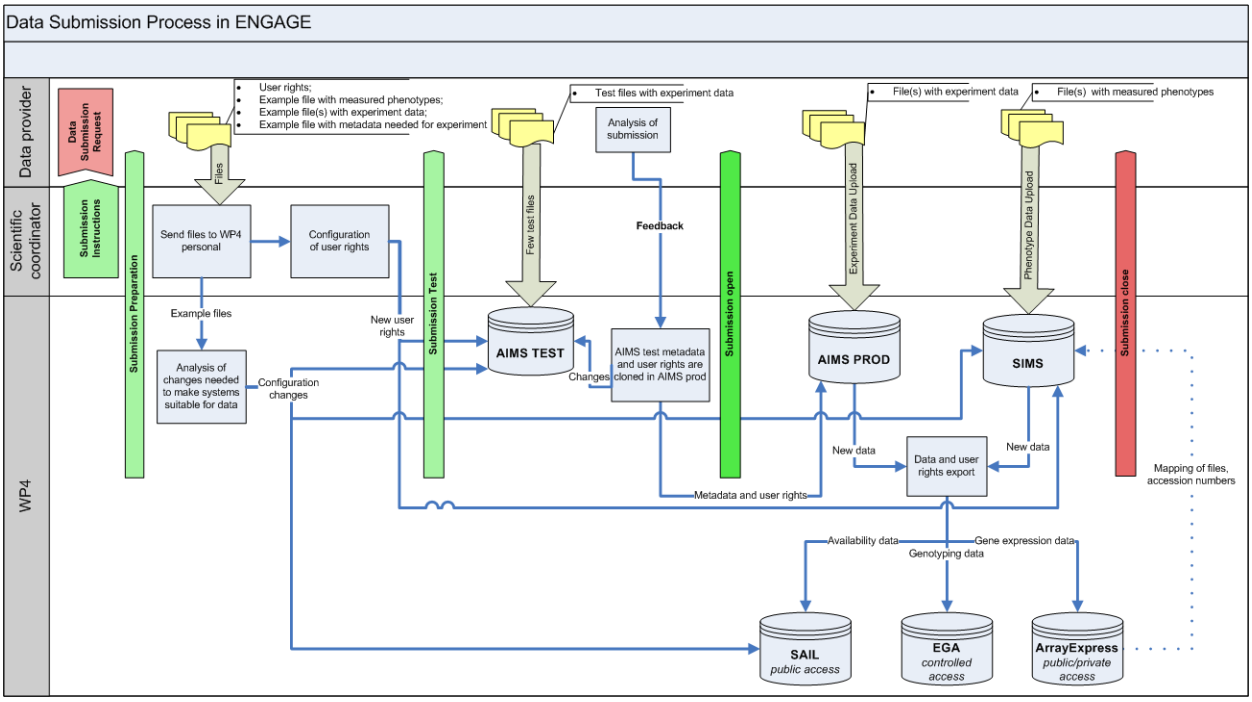
The EGG (Early Growth Genetics) Consortium represents a collaborative effort to combine data from multiple genome-wide association studies (GWAS) in order to identify additional human genome loci that have an impact on a variety of traits related to early growth.

Data access

Summary data files are available at a git repository:
http://www.simbioms.org/gitweb/?p=visualisation.git;a=summary

References

Ferry, G. (2013). Scientific heritage: Science today, history tomorrow. *Nature 493*, 19-21.



Supplementary Figure 1: Engage Data Submission process.