H2020 EINFRA-5-2015

www.bioexcel.eu

Project Number 675728

D3.1 – Selection and Establishment of User Groups

WP3: Consultancy & User Groups

Document Information

Deliverable Number	D3.1
Deliverable Name	Selection and Establishment of User Groups
Due Date	2016-03-31 (PM5)
Deliverable Lead	UEDIN
Authors	Adam Carter (EPCC, UEDIN), Ian Harrow (IHC), Rossen Apostolov (KTH), Stian Soiland-Reyes (UNIMAN), Bert de Groot (GWDG, MPG), Mark Abraham (KTH)
Keywords	User Groups, User Communities, Interest Groups
WP	WP3
Nature	Report
Dissemination Level	Public
Final Version Date	2016-03-23
Reviewed by	Erwin Laure (KTH), Cath Brooksbank (EBI), Steven Newhouse (EBI), Josep Lluis Gelpi (BSC), Berk Hess (KTH)
MGT Board Approval	2016-03-30

Document History

Partner	Date	Comments	Version
UEDIN	2016-02-01	First draft	0.1
UEDIN	2016-03-04	Import of content from Google Docs. Most sections written.	0.2
UEDIN	2016-03-07	Updated with contribution to §4.1.3. Added exec summary.	0.3
UEDIN	2016-03-22	Updated in response to reviewers' comments. Added section on Discourse. Fixed footnote formatting.	0.4
UEDIN	2016-03-23	Updates in response to further reviewers' comments.	0.6

Executive Summary

This document describes the work undertaken in the first five months of the BioExcel project relating to the selection and establishment of user groups.

BioExcel has identified and classified a list of communities of interest and from this a set of initial Interest Groups (IGs) has been selected. This selection is part of an ongoing process of engagement, which is described later in the document. The process of establishing these initial IGs is now underway.

The initial IGs are "Biomolecular simulations entry level users", "Integrative modelling", "Free energy calculations", "Hybrid methods for biomolecular systems", "Best practices for performance tuning", and "Practical applications for industry". A key feature of most of these groups is that they can be built from established user-communities of the project's key codes.

IGs will be supported with mailing lists, and forums, the latter being provided by the Discourse platform. Additional tools and activities such as blogs and real-time chat will be provided in response to the needs of the groups. Webinars will also form an important part of the user engagement process and GoToWebinar has been selected as the platform for this function. These webinars will focus on the interests of the IGs, with an initial webinar planned to highlight BioExcel's offering and to catch the interest of potential IG members.

It is envisioned that BioExcel can function as a hub of IGs (comprised of named individuals who have chosen to be members of the IG) drawn from communities that BioExcel will support. Communities have been classified as "[those] based around research infrastructures and services", "users of particular pieces of software", "users of a particular technique", "application area communities", "industry alliances", "academic initiatives and projects" and "other infrastructure providers".

BioExcel has conducted an initial user survey and has used the results in conjunction with findings from recent user surveys of the project's key codes to inform plans for the initial IGs.

The process for engagement can be summarised as follows: As a first step, members of external communities are invited to join a Discourse support forum and the announcement mailing lists. Initially there will be three categories for each of the pilot codes (Gromacs, Haddock and CPMD); others such as “Automation Workflows” can be added later. With the initial invitation, members of those external communities will be invited to attend the first webinar, which will give an overview of the CoE, its purpose and its support structures. At the first webinar we will announce the suggested themes for future webinars around the five initial IGs. Proposed discussion topics for future webinars will be collected on the forums. Follow-up webinars will be run monthly. If the initial IGs prove to generate activity, new categories will be created for them in the forums. As the IGs gain traction, we will select an “ambassador” from each group who will represent the group in the CoE’s scientific advisory board (SAB).

WP3’s work is ongoing, and further IGs will be established during the course of the project.

Contents

1 INTRODUCTION 6
2 COMMUNITY LANDSCAPING 7
2.1 COMMUNITIES AND INTEREST GROUPS: WORKING DEFINITIONS 7
2.2 COMMUNITIES BASED AROUND RESEARCH INFRASTRUCTURES AND SERVICES 8
2.3 USERS OF PARTICULAR PIECES OF SOFTWARE 11
2.4 USERS OF A PARTICULAR TECHNIQUE 12
2.5 APPLICATION AREA COMMUNITIES 12
2.6 INDUSTRY ALLIANCES 13
2.7 ACADEMIC INITIATIVES AND PROJECTS 14
2.8 OTHER INFRASTRUCTURE PROVIDERS 17
3 USER SURVEYS 18
3.1 INITIAL BIOEXCEL SURVEY 18
4 INTEREST GROUPS (IGS) 22
4.1 INITIAL INTEREST GROUPS 22
4.1.1 BIOMOLECULAR SIMULATIONS FOR ENTRY LEVEL USERS IG 23
4.1.2 INTEGRATIVE MODELLING IG 23
4.1.3 FREE ENERGY CALCULATIONS IG 23
4.1.4 HYBRID METHODS FOR BIOMOLECULAR SYSTEMS IG 24
4.1.5 BEST PRACTICES FOR PERFORMANCE TUNING IG 24
4.1.6 PRACTICAL APPLICATIONS FOR INDUSTRY IG 24
5 TOOLS FOR SUPPORT & SHARING 24
5.1 MAILING LISTS 25
5.2 COMMUNITY SUPPORT FORUMS 25
5.3 WEBINARS 26
5.4 ASK ME ANYTHING 26
5.5 BLOGS 26
5.6 REAL-TIME CHAT 26
5.7 WIKI 26
6 PROCESS FOR ENGAGEMENT 27
7 FUTURE PLANS 28

1 Introduction

To establish a Centre of Excellence, we need to engage with a wide variety of relevant communities and the individuals and organisations that comprise them. As part of this engagement process, BioExcel will need to gain an understanding of the interests, priorities and concerns of the various communities which the centre will support. BioExcel WP3 is involved with both the engagement of users (and potential users) of the centre and also with determining the best mechanisms for offering sustainable services to these communities in the longer term. This process begins with understanding relevant target communities and who the centre could work with in the longer term. It is envisioned that the formation of relevant user groups, or as we have chosen to name them *interest groups (IGs)*, will form an important part of this process of building a community around BioExcel. Thereby ensuring that it brings additional value to the wider community of Computational Biomolecular Research.

In this document we will describe the work undertaken in the first five months of the BioExcel project relating to the selection and establishment of the interest groups. Establishing and engaging with these groups is an ongoing process. Task 3.2 ("Establishment of set of user groups") runs from project month (PM) 4 to PM30 and Task 3.3 ("Active user group engagement") is set to run from PM7 to PM36. Clearly there remains much work to be done in this area, but we have made some good initial progress. Whilst there are no formal updates to this deliverable planned, the work with communities and interest groups is expected to have considerable input to the consultancy proposals. These are expected to form the basis for many of the future deliverables in WP3. Therefore, we plan to provide a short update on interest group establishment and engagement in deliverables at PM18, 24 and 36.

This document begins with a description of community landscaping work (undertaken as part of T3.1) which forms the starting point for establishing interest groups. The Community Landscaping section provides our definitions of communities and interest groups and explains how they are related to each other before going on to describe communities which are considered to be most relevant and important to BioExcel. The User Surveys section describes another aspect to the exploration of the current environment into which BioExcel plans to work. The Interest Groups section describes in more detail about how we are moving from existing communities to form interest groups. We explain how we will establish the initial interest groups. In Section 5, we describe what we as a centre have to offer to members of interest groups. This section explains the benefits we hope to bring to Interest Group members and also the tools that we can offer to help foster communication and collaboration amongst the groups' members and the communities from which they are drawn. In the Process for Engagement section we describe our approach to engagement with communities and the establishment of relevant interest groups.

2 Community Landscaping

There is much activity already ongoing throughout Europe in many of the areas in which BioExcel intends to work. BioExcel aspires to be the pan-European centre around which people working in computational biomolecular research can cluster and associate. Such researchers will already belong to many different communities and organisations. Our aim in undertaking this community landscaping exercise was to gain an initial understanding of how the work of BioExcel relates to these existing communities, and to establish the best ways to support them through the Centre of Excellence.

We start with some definitions to clarify our understanding of the terms we will use in this document and in the project as a whole, before going on to list and classify communities that are most relevant to BioExcel. Whilst we hope that this list is fairly comprehensive, we know that it can never be complete. There will be some communities which we are not yet aware of and there will always be new communities forming, to reflect new science, new techniques, new software, new infrastructures and new interoperability services.

2.1 Communities and Interest Groups: Working Definitions

We consider a **community** to be a group of people who share some common attributes. Many communities will have a common cause or common goal. We therefore define a given community by the attributes that the members of the community have in common. A community will often have some kind of self-organisation, interaction, collaboration and a feeling by the members of the community that they "belong" to the community, but for the purposes of this project we use the term fairly broadly and do not require that all of our communities satisfy all of these requirements.

We consider an **interest group** to be a group of individuals (typically a subset of a particular community) who have chosen to be members of the interest group. The interest groups would be formed by a cross-section of the community members. BioExcel's goal with these interest groups is twofold: The main purpose is to facilitate communication and collaboration *amongst* the people in the group around a clearly defined subject. The second purpose is to maintain open channels of communication *between* the centre of excellence and the various communities that we aim to support, with the IGs taking on the role of representatives of the wider communities.

Thus we envision BioExcel to function as a **Hub of Interest Groups** within the wider computational biomolecular research community.

We have identified and categorized a selection of **target communities** which are considered for recruitment to the BioExcel Interest Groups. These communities can be further subdivided according to other factors such as geographical location, level of expertise or level of experience. Some people may identify themselves as "new users" or "power users" and it might be that this subdivision could be used along with one of the communities below to establish an interest group.

2.2 Communities based around research infrastructures and services

These communities have in common an interest in ensuring that a service/infrastructure remains available and accessible, and have a desire for improvements to the system/infrastructure and compatibility with the system/infrastructure. Many of these communities are larger organisations or projects which may not be directly reachable as a whole, but which can be approached at an institutional and personal level, for instance BioExcel includes three partners who are also involved in ELIXIR.

One important source of possible communities of interest to BioExcel is the ESFRI roadmap, the most recent version of which was published in 2010[1]. At the time of writing, a new roadmap is due to be launched in the next few weeks, and BioExcel will ensure that it explores links with all of the relevant infrastructures in the 2016 update. The table below includes these infrastructures along with other important cross-disciplinary infrastructures such as PRACE.

PRACE http://www.prace-ri.eu	European Research Infrastructure centred on six world-class supercomputers. People using PRACE are already using top-tier high performance computing resources. This community could be supported by ensuring that key biomolecular software is available on, and performs well on PRACE infrastructure. Sub-communities include: **PRACE users** (Service/Infrastructure Users) **PRACE project partners** (Project, Service Provision)
INSTRUCT https://www.structuralbiology.eu	Pan-European research infrastructure in **structural biology**, making high-end technologies and methods available to users. Amongst other things, *Instruct* offers integrated access to multiple platforms, both experimental and computational. This community could be supported, for example, by improving access to, and ease-of-use of the software in the CCISB catalogue. Sub-communities include: **Instruct users** (Service/Infrastructure Users) **Instruct partners** (Service Provision)
EGI http://www.egi.eu/	Federation of computer systems which provide "a seamless grid of academic private **clouds** and **virtualised resources**, built around open standards and focusing on the requirements of the scientific community". EGI could be supported by ensuring that

[1] ESFRI, Strategy Report on Research Infrastructures - Roadmap 2010 (2010). doi:10.2777/23127.

	the BioExcel-supported software and workflow systems are compatible with EGI resources and infrastructure. The nature of grid technologies is such that the separation between users and service-providers is less distinct than, e.g., PRACE, but it is still possible to identify sub-communities: **EGI Users** (Service/Infrastructure Users) **EGI Providers** (Service/Infrastructure Providers)
EUDAT www.eudat.eu	Collaborative pan-European infrastructure for **research data** services, training and consultancy. Focus includes data sharing, data repositories, data transfer and data discovery. EUDAT communities include biomedical and life sciences, earth sciences & environment research, physical sciences & engineering, social sciences & humanities. BioExcel could work with EUDAT to define and promote best-practice in data management and to involve the biomedical research community in EUDAT's work.
ELIXIR www.elixir-europe.org	Pan-European research **infrastructure for biological information**. 15 European countries (and EMBL-EBI) are represented as ELIXIR Nodes, which provide the services and resources that form ELIXIR, including data resources; bio-compute centres; services for the integration of data, software, tools and resources, training and standards expertise. ELIXIR is involved with two large H2020 projects (Excelerate and CORBEL) with a particular focus on biological data resources. BioExcel could offer services which complement the services offered by ELIXIR supporting, for example, HPC applications and workflows.
EU-OPENSCREEN www.eu-openscreen.eu	EU-OPENSCREEN is a collaboration which is "building a sustainable European infrastructure for Chemical Biology, supporting life science research and its translation to medicine, agriculture, bioindustries and society." It is currently seeking ERIC status.
EURO-BIOIMAGING www.eurobioimaging.eu	Euro-BioImaging is a large-scale pan-European research infrastructure project on the ESFRI Roadmap. It will provide open access to biomedical imaging facilities as well as associated support and training. BioExcel's work in workflows and high throughput analysis could be of interest to this community.
WeNMR www.wenmr.eu	Virtual research community focused around NMR and SAXS. The WeNMR portals offers user-friendly access to a variety of services for **structural biology**, some of which are running on the EGI grid. This former FP7 e-Infrastructure project is continuing under the *West-Life* virtual research environment project (see below).

West-Life www.west-life.eu	*West-Life* will pilot an infrastructure for **storing and processing data** that supports the growing use of combined techniques in structural biology. There are some technique-specific pipelines for data analysis and structure determination but little is available in terms of automated pipelines to handle integrated datasets. Integrated management of structural biology data from different techniques is lacking altogether. West-Life will integrate the data management facilities and services (e.g. from WeNMR) that already exist, and enable the provision of new ones. The resulting integration will provide users with an overview of the experiments performed at the different research infrastructures visited, and links to the different data stores. It will extend existing facilities for processing this data. As processing is performed, it will automatically capture metadata reflecting the history of the project. The effort will use existing metadata standards, and integrate with them new domain-specific metadata terms. West-Life will provide the application level services specific to uses cases in structural biology, enabling structural biologists to get the benefit of the generic services developed by EUDAT and the EGI.
BBMRI http://bbmri-eric.eu	An ERIC[2] focused on biobanking and biomolecular resources. This infrastructure does not aim to support biomolecular modelling or simulation, but it is possible that scientists using this infrastructure may also have interests in BioExcel's work. It is possible that BioExcel could compliment BBMRI's work especially at the boundary between computational and experimental biomolecular research. There are also possible relationships between BioExcel's Pilot Use Case 1 (related to high-throughput sequencing analysis in genomics) and BBMRI's involvement in EGI-Engage (http://bbmri-eric.eu/egi-engage) as both are related to high-throughput analysis of experimental results.
ISBE http://project.isbe.eu	*Infrastructure for Systems Biology Europe*. Much of structural biology is not directly in-scope for BioExcel, but the community has been included here because it is a multi-disciplinary community with an interest in modelling. It is possible that the general ideas of biological modelling and how implementations of these models can be worked with in automated workflows could also be of significant interest to parts of the community of users of ISBE.

Communities associated with other ESFRI infrastructures (ECRIN, AnaEE, EATRIS, EMBRC, ERINHA, INFRAFRONTIER, MIRRI) are not considered to be

[2] An ERIC is a European Research Infrastructure Consortium, a legal entity established to operate an international research infrastructure. See https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric.

targets for the BioExcel project. There are possibly lessons that could be learned in terms of sustainability, but these communities are generally expected to be less likely to be important users of a CoE in computational biomolecular research since they are either much more experimentally focused, or they are concerned with biology at larger length scales.

2.3 Users of particular pieces of software

These are users of both "traditional" software which is installed locally as well as Software-as-a-Service. In this category we include software and portals that are maintained or used by BioExcel partners. These communities typically already have mailing lists or forums making them easy to target as a whole.

GROMACS www.gromacs.org	Fast and efficient engine for **Molecular Dynamics** simulations. A BioExcel core application.
HADDOCK www.haddocking.org	Information-driven flexible **docking** software for integrative modelling of biomolecular complexes. A BioExcel core application.
CPMD www.cpmd.org	Parallelized plane wave / pseudopotential implementation of Density Functional Theory compute engine, particularly designed for **ab-initio molecular dynamics.** A BioExcel core application.
Helix Nebula http://helix-nebula.eu/	Partnership between industry, space and science to establish a dynamic ecosystem, benefiting from open **cloud** services for the seamless integration of science into a business environment.
MDWeb and other IRB portals http://mmb.irbbarcelona.org/MDWeb/	Web-based workspace providing standard protocols to prepare **structures**, run standard molecular dynamics simulations and to analyse trajectories
Open PHACTS https://www.openphacts.org	Brings together **pharmacological linked data** resources in a **semantically integrated, interoperable infrastructure**. REST API, workflows, explorer tool, Virtual Machine and Docker implementations. "Doing data plumbing really well!"
Apache Taverna http://taverna.incubator.apache.org/	Open source and domain-independent Workflow Management System – a suite of tools used to design and execute **scientific workflows** and aid in silico experimentation. Dataflow approach.
Kepler https://kepler-project.org	Open source, scientific workflow application. Popular in **physics**. Process-centric, multiple workflow execution

	models.
KNIME https://www.knime.org	Open source, scientific workflow application. Popular in **chemistry**. Dataflow table-driven approach.
Galaxy https://galaxyproject.org	Open source, web-based scientific workflow application. Popular in bioinformatics, particularly for **next-gen sequencing**.
COMPSs http://compss.bsc.es	Open source programming model that allows to program workflows as sequential applications, **automatically parallelising** their execution on distributed infrastructures.

2.4 Users of a particular technique

It could be that the same technique is implemented in different pieces of software, or the technique could require the use of multiple pieces of software. These communities typically do not have any common collaboration mechanism beyond academic conferences and journals.

We can consider, for example, the broad **biomolecular modelling** community, or smaller more focused ones within this, such as **homology modelling**, **docking**, **molecular simulation** or **hybrid QM/MM simulation**.

These communities are possibly best approached through the codes which implement these techniques. It should be possible to build interest groups of users of these techniques which may use different pieces of software and these could provide a forum for comparing, contrasting and sharing the best aspects of different pieces of software.

2.5 Application area communities

These communities consist of people who are working in the same application area. They may be doing different things, but they are working as part of the same industry or research area.

They are similar in some respects to infrastructure communities above, but they are probably sufficiently different to consider separately and possibly target separately. People in one application area might use different infrastructure (local vs international, for example) and several of the infrastructures in the section above are multi-disciplinary. Furthermore, it might be useful to target members of a particular application community that are yet to have gained exposure to a certain infrastructure.

The techniques used to engage with application area communities are likely to be different from those where the community is associated with a code or infrastructure. With application area communities, their benefit from engaging

with BioExcel could be to see what tools, techniques and infrastructures are available to them to help them work in their area.

Important application areas for BioExcel are **Health & Medical**, **Drug Development**, **Biotechnology**, **Environment** and **Agriculture**.

As well as these broad application areas we have related research fields which are applying (or could benefit from applying) computational biomolecular techniques: **biomarker design**, **nanotechnology** and **materials science / material design**, **personalised medicine**, **physiology**, **neuroinformatics** and **chemical modelling** (in particular **multi-scale QM/MM modelling**).

2.6 Industry alliances

Members of these communities belong to an alliance, consortium, partnership or forum in the life science industry. Many of these have existing communities have common interests, purpose and goals. Here we give selected examples to illustrate the enormous breadth of such communities in the life science industry. Some of these examples have particular relevance to BioExcel.

Pistoia Alliance http://www.pistoiaalliance.org	This is a pre-competitive, industry alliance of members, and is represented in the project through the partner IHC. For BioExcel, it is an important target community for building interest in the industrial interest group. The Pistoia Alliance has its own interest groups but none of these currently focus on Biomolecular Research, despite there being many members of this alliance who are active in this area. This presents us with an opportunity to fill this gap.
Open PHACTS http://www.openphacts.org	The Open PHACTS Discovery Platform has been developed to reduce barriers to drug discovery in industry, academia and for small businesses. It is a pilot use case for BioExcel. Originally funded by IMI which transitioned recently to a foundation charity. A BioExcel pilot case.
EMBL-EBI Industry Partnerships http://www.ebi.ac.uk/industry	Quarterly meetings and regular workshops form a loose pre-competitive community. Partners are large life science companies. EMBL-EBI is a BioExcel partner.
EMBL-EBI SME Forum http://www.ebi.ac.uk/industry/sme-forum	An annual forum of SME companies in life sciences with support by One Nucleus (http://www.onenucleus.org). EMBL-EBI is a BioExcel partner.
The European Hit Factory	IMI-funded project that aims to create new chemistry based on crowd-sourced ideas and boost applicants'

https://www.europeanleadfactory.eu	drug discovery programmes at no upfront costs.
European Chemical Industry Council http://www.cefic.org	Cefic is the forum and the voice of the chemical industry in Europe. It is a committed partner to EU policymakers, facilitating dialogue with industry and sharing our broad-based expertise.
FoodDrinkEurope http://www.fooddrinkeurope.eu	FoodDrinkEurope's mission is to facilitate the development of an environment in which all European food and drink companies, whatever their size, can meet the needs of consumers and society, while competing effectively for sustainable growth.
Center for Translational Molecular Medicine http://www.ctmm.nl	CTMM is a public-private partnership for translational research. It is dedicated to the development of medical technologies that enable the design of new and personalized treatments and the rapid translation of these treatments to the patient.
European Biotechnology Network http://www.european-biotechnology.net/	The EBN is a membership organisation whose goal is to improve cooperation in the fields of biotechnology and the other life sciences.

2.7 Academic initiatives and projects

This broad classification includes a variety of academic initiatives and projects. These communities include various groupings of people who are collaborating on specific projects, working together to promote best practices, or sharing infrastructure. The table below includes important international projects but also includes some examples of national-level projects and initiatives. In the latter case, it is harder to make this list comprehensive, but we have tried to identify a selection of established projects and centres in fields related or relevant to BioExcel. This could be used to encourage participation in BioExcel's interest groups.

SBGrid consortium https://sbgrid.org	Support structural biologists by providing structural biology laboratories with a tested and refined software infrastructure that includes a large library of scientific applications. Members also benefit from access to SBGrid-supported high performance computing (HPC) resources and training opportunities. This consortium is US-based and could provide opportunities for collaboration and sharing of best practice beyond Europe.
CCPBioSim project	Collaborative Computational Project for Biomolecular

http://www.ccpbiosim.ac.uk	Simulation, UK. This project is well-established and the obvious possible offering from BioExcel here (and with HECBioSim below) is to expand the existing UK biomolecular simulation communities and promote interaction with the wider community across Europe.
HECBioSim project http://www.hecbiosim.ac.uk	This project promotes the use of High-End Computing Resources by the Biomolecular Simulation Community in the UK and it is closely related to CCPBioSim above, having a particular focus on HPC simulation.
CCP5 http://www.ccp5.ac.uk	Another UK Collaborative Computational Project, CCP5 focuses on computer simulation of condensed phases. There is overlap here with BioExcel's work, and again BioExcel could bridge the gap between CCP5 members and researchers elsewhere in Europe working in this area.
SSI http://software.ac.uk/	Software Sustainability Institute, UK. Experts in software engineering good practice especially for research software. BioExcel could participate in (two-way) sharing of best practice.
DEEP-ER http://www.deep-project.eu	Exascale projects. These projects are working at the longer-term and high-performance end of the spectrum and so it is likely that our interests will not closely align at first, however it could be that expert HPC users in BioExcel could be interested in interacting with these projects to help shape future HPC systems. BioExcel could provide a bridge between the work of these projects and the Biomolecular research community. DEEP-ER currently has no biomolecular co-design applications. MontBlanc has BUDE[3], which is a molecular docking program, so there is potential here for sharing of ideas with HADDOCK.
MontBlanc https://www.montblanc-project.eu	
E-CAM http://www.e-cam2020.eu	Another H2020 Centre of Excellence. This centre is focused on simulation and modelling. There is no specific application targeted here (unlike BioExcel) but it could be that communities which are not application area-focused (such as experts in developing certain codes) could participate in BioExcel's activities and benefit from our future services.
MaX http://www.max-center.eu	Another H2020 Centre of Excellence. This centre is focused on Materials Design (at Exascale). Here the most likely overlap again is with experts in certain techniques and there are possible joint interests here in terms of

[3] http://www.bris.ac.uk/biochemistry/research/bude

	workflows and automated high-throughput calculations in what they refer to as "the data and applications ecosystem"[4]
Dutch TechCenter for Life Sciences www.dtls.nl	DTL offers facilities, services, courses and community services in many areas of life sciences across the Netherlands. This organisation could act as a bridge between Dutch life sciences communities and wider communities across Europe.
BioValley http://www.biovalley.com/academia/excellence-in-research	An initiative supporting Life Sciences (both industry and academia) in parts of France, Germany and Switzerland. BioValley offers matchmaking, promotion and networking services, but they do not obviously offer computational expertise, which could potentially be offered by BioExcel.
BluePrint http://www.blueprint-epigenome.eu	A large-scale research project working in epigenomics. With 39 partners working in what is considered to be a high-impact project, this could be a source of potential members of interest groups. The project works in several areas including high-throughput analysis.
Swedish e-Science Research Centre (SeRC) http://e-science.se	National initiative between the largest universities and HPC centres in Sweden to support the e-Science communities. Possible link between Swedish and wider European communities.
eSSENCE http://essenceofescience.se	A strategic collaborative research programme in e-science in Sweden. Possible link between Swedish and wider European communities.
Bioinformatics Infrastructure for Life Sciences (BILS) http://bils.se	Providers of structure prediction servers Pcons[5], PconsC[6], TOPCONS[7], SCAMPI[8]. These are deployed in the EGI Federated Cloud using vo.nbis.se. The community of users of these servers could possibly benefit from workflows and services being developed by BioExcel. There is also possible best-practice that could be shared with regard to integration with the EGI cloud.

[4] http://www.max-center.eu/ecosystem_and_data/
[5] http://pcons.net
[6] http://c.pcons.net
[7] http://topcons.cbr.su.se
[8] http://scampi.cbr.su.se

Performance Optimization and Productivity (POP) https://pop-coe.eu	POP is another Centre of Excellence, which offers software optimisation services. POP is coordinated by BSC, a BioExcel partner, and has a potential interaction for the improvement of core-codes in the project.
Multi-scale complex Genomics (MuG) http://www.multiscalegenomics.eu	MuG is a H2020-funded Virtual Research Environment focused on 3D/4D genomics. MuG has a special interest in biomolecular simulations from nucleic acids to chromatin levels, what aligns directly with BioExcel aims. MuG shares several partners (EBI-EMBL, IRB, BSC) with BioExcel.

2.8 Other Infrastructure Providers

This class of communities includes organisations like **cloud service providers** and **independent software vendors** (ISVs). As well as large cloud providers (such as Amazon[9], Microsoft[10] and Google[11]) there are a growing number of smaller cloud service providers offering hardware-as-a-service, platform-as-a-service and software-as-a-service. The last of these is a growing area and there is a blurring between software and services in many cases with online services offering both browser-based software and APIs that can be accessed by programs running elsewhere. Whereas pipeline and workflow software used to be considered fairly advanced and specialised, there are a growing number of consumer offerings (e.g. IFTTT[12], Workflow[13]) for the automation of tasks that themselves run as software in the cloud. Whilst these consumer services are less likely to be of use to researchers, it means that that people are more familiar with the idea of online computing and more effort is being put into software which can be used in this way. The increasing interoperability and wider availability of web APIs means that it is becoming possible to run scientific workflows with components being offered by multiple infrastructure and service providers.

As well as the communities of *providers* there are also various communities of users of these services. Many users -- Amazon Web Services uses, for example -- may not see themselves as a community, but they do have common interests, both things that might have drawn them to a large cloud-compute provider in the first place, and an interest in having other things that work in the Amazon cloud, so that they can work in a familiar environment and have software and services interoperate smoothly. In terms of an offering to these user communities,

[9] https://aws.amazon.com
[10] https://azure.microsoft.com
[11] https://cloud.google.com
[12] https://ifttt.com
[13] https://workflow.is

BioExcel can keep in mind interoperability with popular infrastructure platforms.

Interoperability services are of vital importance for the Open PHACTs platform[14] which is founded on semantic web and linked data principles and uses industrial strength tools such as Virtuoso[15] to provide fast and robust access to chemistry and biological data sources. Users can access the data via the Open PHACTS API or explore it using the Open PHACTS Explorer and many other apps developed using the API.

3 User Surveys

As part of the ongoing engagement with interest groups, we plan to undertake several surveys during the course of the project. In order to help plan these surveys and also to gain some basic insight into the requirements and interests of potential centre users we conducted an initial survey amongst contacts of the project's partners. We also looked at results from recent surveys of the users of HADDOCK and GROMACS communities. Results from these are discussed below.

3.1 Initial BioExcel Survey

The initial BioExcel survey was sent out to a number of colleagues and contacts of the partners in the BioExcel project. Since it is relatively early in the project, this activity took place in parallel with identifying target communities and before the initial IGs were determined (indeed, feedback from the surveys has helped to inform the decision as to what IGs we could initially establish). For this reason, the number of people who were invited to complete the survey is small. Since the project partners include code developers and centres at which the software is used it provided a means for us to obtain useful insight, but it is recognised that wider surveys will be necessary in the course of the project to gain more representative results. Future surveys will be sent to much larger numbers of people, but we did not want potential users of the Centre of Excellence to feel like they were being spammed or asked questions that might be less relevant to them before the real process of Interest Group engagement begins in earnest[16].

The survey had been sent to 50 people and 24 completed submissions were available at the time of analysis[17]. In some cases, we have tried to separate out responses from people who are actively using the software and people who are not, but the small numbers of responses preclude quantitative analysis at this level of detail.

---

[14] https://www.openphacts.org
[15] http://virtuoso.openlinksw.com
[16] Community engagement is, in fact, already underway, but it should be noted that the engagement task described in the project's Description of Action is not due to start until project month 7.
[17] At the time of writing, the survey remains open in order to gather further input.

Responses to the survey were, on the whole, quite varied although there were some important biases that should be borne in mind when interpreting the results: 18 respondents (75%) are working in structural biology and 19 respondents (79%) are looking at proteins. 17 respondents (71%) are studying dynamics.

In terms of codes being used: 46% are currently using GROMACS, 21% are currently using HADDOCK. 0% are currently using CPMD and 79% had never tried using CPMD. This last point is not surprising: The barrier to entry to biomolecular CPMD is unusually high because you need to have a suitable problem, access to a big enough computer, get a CPMD license, and (in the case of QM/MM) pay for a GROMOS license[18]. BioExcel's work is to lower this barrier of adoption by interfacing CPMD with the free GROMACS as the MM engine in place of the proprietary GROMOS. In addition, GROMACS is much faster and versatile. Thus we hope to be able to establish a user community for biomolecular QM/MM with CPMD by the end of the project.

A variety of other codes and platforms were mentioned as being used by the respondents, namely Amber[19], NAMD[20], CHARMM[21], CP2K[22], OCCAM[23], CNS[24], Xplor[25], CYANA[26], MOE[27], firefly[28], Gaussian[29], molpro[30], gamess-us[31], Turbomole[32], OpenMM[33], ACEMD[34], Plumed plugin[35], Desmond[36], Rosetta[37] and Coot[38].

---

[18] A brief online search found almost no papers using QM/MM with CPMD+GROMOS.
[19] http://ambermd.org
[20] http://www.ks.uiuc.edu/Research/namd/
[21] http://www.charmm.org
[22] https://www.cp2k.org
[23] https://www.molnac.unisa.it/occam/project.html
[24] http://cns-online.org/v1.3/
[25] http://nmr.cit.nih.gov/xplor-nih/
[26] http://www.cyana.org/wiki/index.php/Main_Page
[27] https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm
[28] http://classic.chem.msu.su/gran/gamess/index.html
[29] http://www.gaussian.com
[30] https://www.molpro.net
[31] http://www.msg.ameslab.gov/gamess/
[32] http://www.turbomole.com
[33] http://openmm.org
[34] https://www.acellera.com/products/molecular-dynamics-software-gpu-acemd/
[35] http://www.plumed.org
[36] http://www.deshawresearch.com/resources_desmond.html
[37] https://www.rosettacommons.org/software
[38] http://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/

Respondents were asked to note their main challenges. Results are summarised in Figure 1. **Parameter tuning** was the most common challenge encountered. This probably relates in some cases to a deeper problem of understanding the system in question, but there are things that could be done to address this. There are probably two main classes of parameters here, ones which only affect performance and ones that could affect correctness. The former is probably easier to address whereas correctness of MD or other software parameters is still an active research topic. Historically, it has been hard to address because of low availability of software that can both implement the model correctly, and do so fast enough to show for reasonable computational cost whether use of the model leads to valid conclusions on a problem size of interest. Some things such as **compilation and installation** (5; 23%) and **configuration and commands** (6; 27%) are possibly low-hanging fruit in terms of things that the CoE could help with.

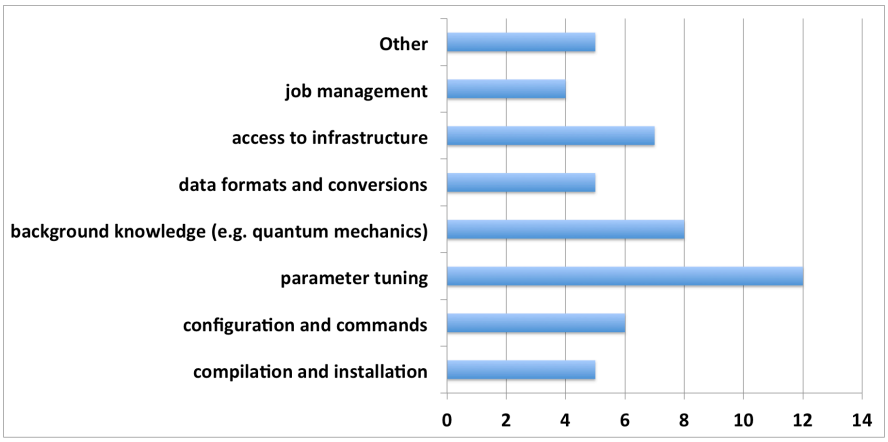

Figure 1: Main Challenges identified in Initial Survey. The x-axis shows the number of respondents.

The survey then asked about tool integration. The results show that integration of molecular modelling tools in larger analytic workflows is mostly done either manually or using scripting (Python being widely mentioned). It should be noted that most of the respondents who did some things manually also used scripting. A relatively small number use scientific workflow systems. Amongst those who do, the following Tools and systems were used: Pipeline Pilot, Python scripts, Bash scripts, Galaxy, KNIME, jBoss DV, Tibco. The most frequently mentioned step which people feel they could benefit from expert support was **Analysing/correlating multiple data sets from calculation(s)**. We believe that the kind of support people are looking for here is primarily scientific (or best practice in statistics/analysis) rather than technical issues of bringing datasets together for analysis, although this point could be clarified in future questionnaires.

Next, we consider access to HPC infrastructure. A variety of different kinds of computers are being used for molecular simulation, with **compute clusters** being most frequently mentioned (17 respondents; 77%). **Supercomputers** rank second (15; 68%), and **desktops** rank third (11; 50%). There are also a significant fraction of users using **Grid** (23%) and **Cloud** (18%) platforms. A possible issue with regard to take-up of cloud platforms is one of security,

particularly for commercial/industrial users. Whilst private (and hybrid) cloud solutions are available, these require more work to set up and maintain. Respondents appear to generally prefer **self-managed computers** (or computers managed within their own research group) (16; 73%) compared with a smaller number who use **national** (11; 50%) and **European** (7; 32%) services. It is expected that ease-of-use and familiarity are the main factors here, but this could be explored in future surveys.

A significant number of respondents (7, 64%) use the (remote) HADDOCK web server, which backs up the assertion that people generally consider remote-access tools to be acceptable for their research (15; 68%). This is further backed up by the HADDOCK user survey in which 92.1% (575 respondents) use the web server version and 34.6% (217 respondents) *only* use the server version. Reasons for not using a remote web service for HPC applications included concerns about the number of jobs, the size of output files, lack of control over input, privacy/intellectual property, difficulties with remote file management, lack of need, lack of trust and lack of reliability. With the possible exception of privacy, it seems that many of these issues could be addressed by BioExcel. The HADDOCK user survey asked those who did not use the web server why this was the case. Answers included: “A preference for a local installation because the process of optimising models is more streamlined”, “It is easier to run parallel jobs” and “Limitations on transfer of data outside own organisation”. The HADDOCK user survey also asks the question why people do not use a local installation: Several respondents had only heard of the online version; more than one respondent stated that the online version was faster; multiple respondents stated that they did not have sufficient access to local compute facilities and more that one respondent stated that they didn’t have expertise in installing the software or that it was complicated to use.

The survey finished with an open question about what else BioExcel should focus on and the answers were varied:

- Tutorials related to different computational examples
- Development of programs (e.g. something “similar to Schrodinger/Desmond Simulation Interaction Diagram to analyse MD simulation data just from one window”, “an iTunes like interface that I can use from my desktop to manage calculations on remote HPC centres, possibly as a frontend to Copernicus”
- Integration of hardware and software platforms (e.g. for genome computing)
- Improvements to documentation (GROMACS was singled out here)
- Improvements of aspects of key software (e.g. “integration of the GROMACS analysis programs so [it] could run with a single configuration script”)
- Improvements to tool-chain for force field parameters calculation
- Integration with pipelines (KNIME, Galaxy)
- Student training on how to appropriately use simulation tools

It is possible that some of these could be addressed in the BioExcel project but other ideas here could form the basis for ideas for future work of the Centre of Excellence. Future surveys could offer a list of suggestions (possibly generalised versions of some of the ideas here) that could be prioritised by respondents.

4 Interest Groups (IGs)

This section describes the initial interest groups, describes the rationale for selecting them, describes how we hope to engage with them and provides certain metrics such as number of users, extent of interaction.

From the long list of communities described above, it is clear that there is a wide choice of possible communities that we could aim to establish user groups from. In choosing the initial interest groups, we have aimed to benefit from the expertise and links that the project partners already have. Since the code owners are actively involved in the project, it makes sense to target IGs that could benefit from the expertise and involvement of the code developers. The centre's three main codes and the various workflow platforms have thousands of users worldwide, and are being used for a diverse set of problems, including non-biological studies. In order to achieve maximum impact for the benefit of the computational biomolecular researchers, we have selected several areas that are narrower in scope than the code's user groups while still sufficiently important to a large subset of the codes' users.

In the longer term, we will aim to learn from our experience in establishing these initial interest groups. The level of interest will help us to gauge whether we should be targeting more precise and specific areas of interest, or whether it makes more sense to have broad interest groups with more participants. As measures of "success" of interest groups there are various metrics that will be considered, including:

- the number of people signed up to the group,
- the number of posts in the forum associated with the group,
- the number of emails sent in the mailing list for a group, and
- the responses from surveys sent to members of the IGs.

4.1 Initial Interest Groups

Initially, the groups will be organised around users of a given core application with the addition of a few separate ones. This aligns with the functions of the centre since our products are the pilot codes and workflow systems, and our services deliver the provisioning of support through applying best practices and training on that software.

At the same time, the topics of the initial IGs are not code-specific, but are rather defined as "horizontal" focus areas. Our expectations are that over time they will attract users of other codes.

We have thus defined the following six interest groups:

4.1.1 Biomolecular Simulations for Entry Level Users IG

This is a group for scientists working on biomolecular (proteins/nucleic acids) structures but with little or no experience of setting up simulations. That would be users who are mostly interested in simulations in standard conditions (constant temperature and pressure, explicit solvent, no ligands) seeking to solve specific biochemical questions (effects of mutations in stability, simple conformational changes) using biomolecular simulations.

The IG targets mainly users of portals for automated simulations such as those offered by IRB: **MDWeb** (http://mmb.irbbarcelona.org/MDWeb), a platform to setup, run and analyse **protein structure** molecular dynamics, **FlexServ** (http://mmb.irbbarcelona.org/FlexServ), to run a complete set of flexibility analyses on protein structure MD trajectories, and **NAFlex** (http://mmb.irbbarcelona.org/NAFlex), a platform to build, setup, run and analyse **nucleic acids** molecular dynamics simulations.

In addition, those IG users will have access to a set of already computed MD simulations stored in a couple of databases: **MoDEL** (http://mmb.irbbarcelona.org/MoDEL), with almost 2000 different protein structures trajectories and **BigNASim** (http://mmb.irbbarcelona/BigNASim), with 150 nucleic acid trajectories covering 150 μs simulation time.

For entry level users interested in small molecule docking, the **SEABED** web server (http://www.bsc.es/SEABED) integrates a variety of docking and QSAR techniques in a user-friendly environment, going beyond the basic docking and QSAR web tools, implementing extended functionalities like receptor preparation, library editing, flexible ensemble docking, hybrid docking/QSAR experiments or virtual screening on protein mutants.

Leader of the group is IRB.

4.1.2 Integrative Modelling IG

This group is for users who have an interest in the structural prediction of biomolecular interactions combining experimental data from various sources with powerful algorithms to generate high-resolution 3D models of the macromolecular complexes. The time has come to 'combine and conquer'.

This IG mainly targets the HADDOCK users. Leader of the group is UU.

4.1.3 Free Energy Calculations IG

This group will discuss topics of interest related to methods for free energy calculations, which are crucial for evaluation of reaction paths, binding propensities (drug design), thermostability predictions, protein design, etc. There will be a particular focus on the automated, unsupervised setup of such calculations, rendering such simulations more user-friendly and accessible to the novice user as well as enabling large-scale scans for advanced users.

This IG mainly targets the Gromacs users. Leader of the group is KTH.

4.1.4 Hybrid Methods for Biomolecular Systems IG

This interest group will deal with hybrid methods in general, in order to better describe biological systems. In particular, the discussions will be concerned about the methodologies and the approaches to interface the different granularity of descriptions from quantum to coarse levels.

The IG targets mainly CPMD users. Leader of the group is JUELICH.

4.1.5 Best Practices for Performance Tuning IG

Discussion and information sharing group for best practice in running efficiently on HPC infrastructure. Scientific quality is often in direct proportion to the amount of computation run, so running efficiently is quite important. There are several auto-tuning tools, publications, and existing general knowledge that can help narrow the search space for users. Identifying important target simulations will also help developers prioritise improvements.

The group will initially target GROMACS users and gradually expand to the other codes. Leader of the group is KTH.

4.1.6 Practical Applications for Industry IG

We are developing a strategy for the formation and engagement of an Industry IG which will place emphasis on the practical applications of computational biomolecular research. This will consider any prior use cases of relevance to industry to show the impact and benefit of biomolecular modelling and simulations. It will determine what aspects generate the most interest and provide support for practical application through understanding the needs and challenges of this IG who are important for expanding usage and developing our sustainability plans.

Leader of the group is IHC.

5 Tools for Support & Sharing

Nucleation and development of new communities is typically a long and arduous process. It requires potential members to easily see a value in participating in the community, and in order to be sustainable, this participation needs to be internally driven. Numerous research on the topic has shown the importance of the following motivation factors:

- recognition among peers
- influence within the group
- sense of belonging to the community

In order to address and stimulate those factors, the centre will offer a variety of tools for information sharing and provision of expert support by the CoE as well as community support by the IGs.

5.1 Mailing lists

Mailing lists are simple and commonly used for community engagement. They are currently the main communication and discussion medium for users of the pilot codes with thousands of subscribers.

However, mailing lists do not excel at providing support to the main motivation factors above. Based on the experience of project partners who have maintained mailing lists for over 15 years, some of their drawbacks are quite apparent:

- no overview of “hot” topics
- no overview of who is contributing most
- no tags/categories etc.
- duplication of answers and questions
- can't update answers as new information comes along
- hard for users to judge the quality of the answer / answerer
- search is not optimal
- younger people don't know the etiquette because mailing lists are becoming rare
- mailing list sign up / unsubscribe is something of a barrier to entry
- people fail at using digests
- people fail at starting new topics

Modern forum based software solutions are able to address all of the above issues and provide a much better and useful user experience. Full description is provided in the following section.

For the needs of BioExcel’s community engagement, we will use IG-dedicated mailing lists only for announcements, and refer members to the community forums for discussions and support.

5.2 Community support forums

The Discourse discussion forum (https://www.discourse.org/) is a project that grew out of its founders' successful experience at StackExchange, Unlike StackExchange, it is not limited to the "question and answer" format, and is intended to be a modern discussion forum that will replace all such existing solutions. This seems like an ideal platform for the BioExcel interest groups.

It provides many benefits, including

- in the common situation where users wish to share attachments showing inputs and outputs, mailing these to thousands of subscribers is anti-social;
- facilitating searching for content, because discussions can be tagged, duplicates recorded, and answers updated as new information comes to hand;
- having voting and reputation systems that make it easier for a user to estimate the quality of the answer or answerer;

- simplifying sign-up procedures through social-media-based single-click login;
- providing a platform for maturation of user-generated material that might evolve into official documentation;
- being designed for readability on mobile platforms;
- removing the motivation for email digests, which fragments mailing-list discussion and archives when people reply to them; and
- providing user number and activity statistics, to help quantify the size and value of the community.

5.3 Webinars

Online webinars are efficient means for the delivery of specific training to a large and distributed audience, as well as engaging with users. The centre is planning to organize monthly webinars to address important aspects about the usage of the three main codes in WP1 (or selected tools/workflows from WP2) that address the needs of the interest groups. The webinar will provide a direct measure of user engagement and relevance of provided material. They will be invaluable for gathering feedback during scheduled Q&A sessions at the end of the events.

5.4 Ask Me Anything

Depending on the generated interest and needs of the IGs, the centre will consider offering regular ask-me-anything (AMA) sessions during which members of the IGs can discuss their problems directly with the core developers and centre experts. The format, duration and scheduling of the sessions will be decided at later stage according to the IGs' needs.

5.5 Blogs

BioExcel can offer to IG members the opportunity to publish blogs about their ongoing research work on the centre's website. This will increase the visibility of the researcher's work and give an opportunity to be noticed by a wider audience.

5.6 Real-time chat

Real-time chat channels are very popular in many organizations and communities. Depending on IG needs, BioExcel can provide custom spaces for discussions such as Gitter rooms (gitter.im) or Slack channels (slack.com).

5.7 Wiki

Wikis are convenient systems for organizing user generated content such as knowledge bases. In newer communities it may be challenging to enforce contributions but if the IGs express interest in sharing and organizing certain content using a wiki, then BioExcel will provide the necessary platform.

6 Process for engagement

Figure 2, below, presents a diagram of the nucleation and engagement of IGs. As a first step, members of external communities are invited to join a Discourse[39] support forum at ask.bioexcel.eu [**1**] and the announcement mailing lists at www.bioexcel.eu [**5**]. Initially there will be three categories for each of the pilot codes (Gromacs, Haddock and CPMD), and others such as "Automation Workflows" can be added later. With the initial invitation, members of those external communities will be invited to attend the first webinar [**2**], which will give an overview of the CoE, its purpose and its support structures. At the first webinar we will announce the suggested themes for future webinars around the five initial IGs [**4**]. Proposed discussion topics for future webinars will be collected in special topics on ask.bioexcel.eu [**3**]. Follow-up webinars will be run monthly. If the initial IGs prove to generate activity, new categories will be created for them in the forums [**6**]. As the IGs gain traction, we will select an "ambassador" from each group who will represent the group in the CoE's scientific advisory board (SAB) [**7**].

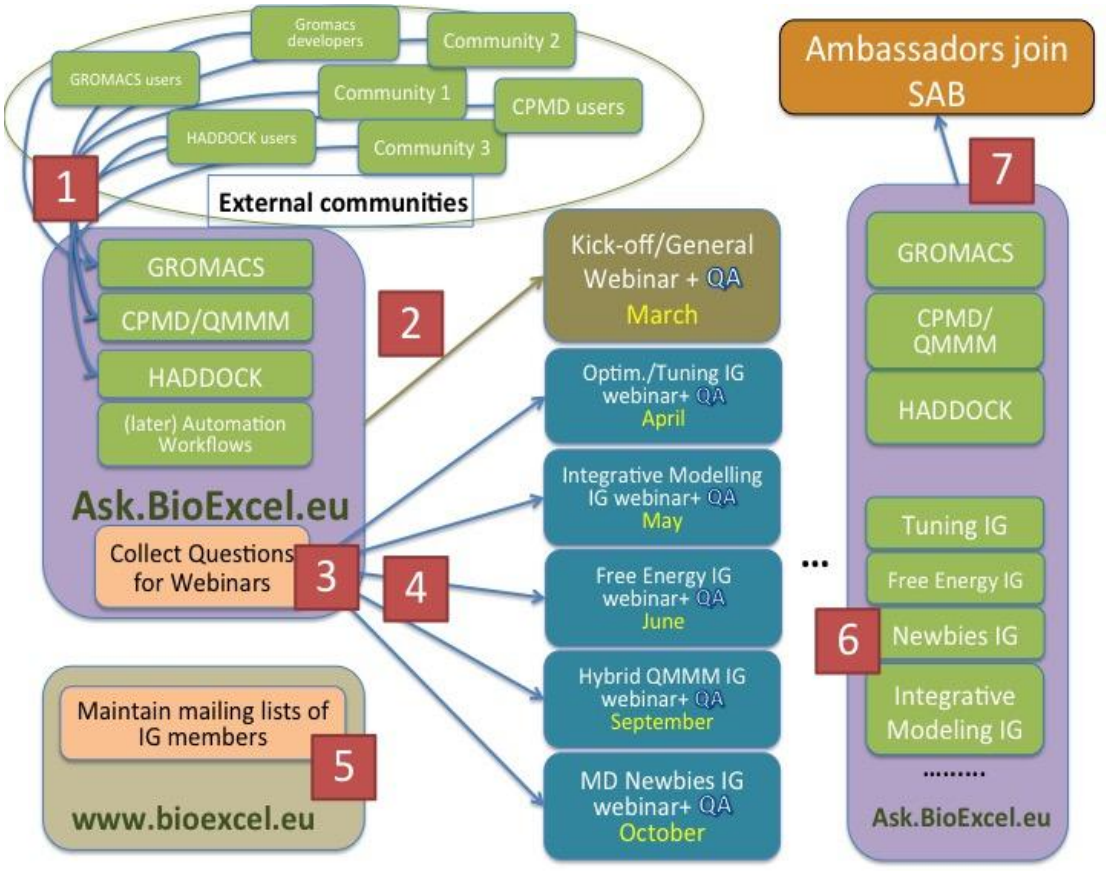

Figure 2: Planned Process for Engagement

[39] http://discourse.org

7 Future Plans

We will continue to establish further interest groups as the project proceeds. The engagement process described above is expected to be sufficient to recruit both initial and future IGs, but we will also investigate semi-automated mechanisms for people to request and set-up their own interest groups.

This process will be improved in future as we learn what works well and where improvements can be made. We expect that much of the future interest will be driven by the frequently asked questions being posed by the community of biomolecular scientists, through our communication channels such as webinars, Discourse, Gitter, GitHub and the wiki. Interest Groups will feed into the original user groups for each core application, which exist already for GROMACS and HADDOCK. A similar group (i.e. outside of BioExcel's IG) is also planned for CPMD users.

We recognise that establishing the IGs is not an end in itself, and that continued effort will be required to engage in the IGs. We expect to have project members (such as the code developers) actively participating in the discussions of the IGs. We also expect to hold face-to-face IG meetings, where appropriate, to strengthen the communities around the IGs. As part of this engagement, we aim to capture the requirements of the IGs and the communities they represent, introduce them to new developments in the applications that the project is working on and gain as much understanding as possible of their workflows so that we can support them now and in the future.

As the project's pilot use-cases become further established, there is also the opportunity to create interest groups related to the work in these use-cases. At the same time, if a given group doesn't generate the expected interest and pick up momentum, we will consider closing it and focusing efforts on a different area.