Using virtual research environments in agro-environmental research

R.M. Lokers[1], M.J.R. Knapen[1], L. Candela[2], S. Hoek[1], W. Meijninger[1],

[1] Wageningen Environmental Research, Droevendaalsesteeg 3, 6700 AA, Wageningen, the Netherlands
[2] Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy, Pisa, Italy

Abstract. Tackling some of the grand global challenges, agro-environmental research has turned more and more into an international venture, where distributed research teams work together to solve complex research questions. Moreover, the interdisciplinary character of these challenges requires that a large diversity of different data sources and information is combined in new, innovative ways. There is a pressing need to support researchers with environments that allow them to efficiently work together and co-develop research. As research is often data-intensive, and big data becomes a common part of a lot of research, such environments should also offer the resources, tools and workflows that allow to process data at scale if needed. Virtual research environments (VRE), which combine working in the Cloud, with collaborative functions and state of the art data science tools, can be a potential solution. In the H2020 AGINFRA+ project, the usability of the VREs has been explored for use cases around agro-climatic modelling. The implemented pilot application for crop growth modelling has successfully shown that VREs can support distributed research teams in co-development, helps them to adopt open science and that the VRE's cloud computing facilities allow large scale modelling applications.

Keywords: Virtual Research Environment, agro-climatic modelling, data science, big data, crop growth modelling

1 Introduction

Agro-environmental research is highly interdisciplinary, and therefore researchers in the field are generally accustomed to linking with and working together with peers over different scientific domains. However, todays research challenges as captured for instance in the Sustainable Development Goals [1] or in the EU's societal challenges [2] require new approaches, that take advantage of data science and open science practices when combining even more cross-sectoral and cross-discipline knowledge, information and data. Lokers et al. [3] state that recent trends like (1) broadening policy and decision contexts for research; (2) the attention to open data as a public good resource; (3) the tremendous growth of the amount of data available for science and (4) the massive increase in computational resources, and better availability and accessibility of

them, as well as of data storage in the Cloud, have greatly increased the opportunities for data science to use big data. In Europe, the EC has recognized these trends and has taken action by establishing the European Open Science Cloud (EOSC) [4], which aims to provide researchers better access to high performance distributed compute and storage resources and to better facilities to share their data and tools with the research community in the large.

At present, a lot of research is still carried out in relatively closed communities, with data and knowledge being used and reused in a limited way, among a fairly small network of trusted peers. In operational (data) science, addressing societal challenges and innovating on a global scale usually requires collaboration between multidisciplinary teams. To advance interdisciplinary science and adopt data science and its many opportunities, researchers from different domains and knowledge networks will have to connect, collaborate and co-develop more intensely and on a larger scale than ever before. This can be achieved through the establishment of collaborative, cloud based working environments that enable remote groups to work together efficiently as a team. Such environments use state-of-the-art ICTs to develop, share and reuse resources and to comply with the requirement to publish and process heterogeneous big data resources. Virtual research environments, offering cloud based facilities for collaboration, social interaction and a range of facilities for performing data science, e.g. data discovery and data sharing, data wrangling, distributed computing facilities and visualization, are aiming at providing such environments. In the H2020 AGINFRA+ project, agri-environmental researchers from different domains, infrastructure experts and software developers work together to co-develop a VRE that supports agro-environmental data science and implements typical agri-food and agri-environmental use cases.

2 Virtual Research Environments

In recent years, many infrastructures, science gateways and VREs have been developed, tested and used in scientific practice. Science gateways, virtual laboratories and virtual research environments are all terms used to refer to community-developed digital environments that are designed to meet a set of needs for a research community [5]. Specifically, they refer to integrated access to research community resources including software, data, collaboration tools, workflows, instrumentation and high-performance computing, usually via Web and mobile applications. Such infrastructures support researchers with a tremendous range of different functions. Ahmed et al. [6] examined a large amount of VREs and found them to be most commonly characterized by ICTs that support communities and that allow posting and transferring information, with tools like Databases, Instruments and Computational ICTs being much less common (but also important for more substantive types of VREs). They also found that community ICTs in these VREs were almost entirely dedicated to providing one-way transmission of information and that interactive tools such as chat systems and conferencing systems are almost never incorporated in VREs. Application of VREs can be found over various thematic areas. Zuiderwijk et al. [7,8] describe multidisciplinary VRE requirements for the use of data coming from governments and publicly-funded research

organizations. They show how meeting these requirements results in a VRE that 1) overlays existing e-Research Infrastructures to provide researchers with integrated open data from different domains, 2) offers Open Government Data in combination with data from publicly-funded research, and 3) stimulates innovation and research collaboration.

In the context of the H2020 AGINFRA+ project it was decided to follow the *system of systems* approach [9] to develop a comprehensive platform enacting to set up and operate several Virtual Research Environments with the as-a-Service delivery model. This platform is implemented by assigning a pivotal role to the D4Science infrastructure and blending together "resources" from "domain agnostic" service providers (e.g. D4Science [10], EGI [11], OpenAIRE [12]) as well as from community-specific ones (e.g. AgroDataCube [13] [], AGROVOC [14], RAKIP model repository [15]) to build a unifying space where the aggregated resources can be exploited via VREs [16]. The resulting platform is depicted in Figure 1, described in [17] and made available through a dedicated gateway[1].

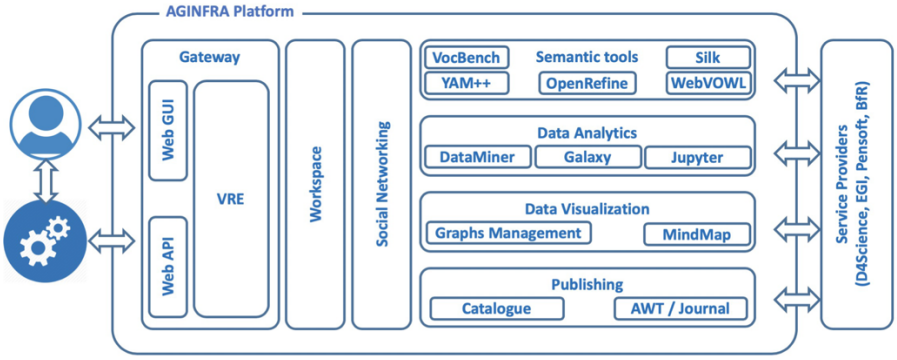

Fig. 1. The AGINFRA+ Platform Architecture

The D4Science is at the heart of the overall platform, offering core services including (*a*) the AGINFRA+ *gateway*, realising the single access point to the rest of the platform; (*b*) the *authentication and authorisation infrastructure*, enabling users to seamlessly access the aggregated services once managed to log in the gateway; (*c*) the *shared workspace*, for storing, organising and sharing any version of a research artefact, including dataset and model implementation; (*d*) the *social networking area* enabling collaborative and open discussions on any topic and disseminating information of interest for the community, e.g. the availability of a research outcome; (*e*) the *overall catalogue* recording the assets worth being published thus to make it possible for others to be informed and make use of these assets. Most importantly, it offers the facilities for setting up and operating Virtual Research Environments.

These basic facilities for virtual research are complemented by services for the semantic-oriented management of data, data analytics, data visualization, and publishing [17].

[1] AGINFRA Gateway https://aginfra.d4science.org/

3 Virtual research for agro-climatic modelling in AGINFRA+

3.1 AGINFRA+ and the agro-climatic user community

AGINFRA+ aims at using VREs to bring open science forward in agri-food research. The AGINFRA+ initiative serves a range of scientific user communities in the agri-food domain and implements and evaluates virtual research supported pilots in a variety of use cases that are relevant for these communities. One of the target communities for AGINFRA+ is the agro-climatic modelling community. This group of researchers focusses on developing and calibrating agro-environmental models and algorithms and applying these in research in the agro-environmental and food security domains. They use a variety of agro-environmental data, for instance agronomic data (like crop parameters, crop calendars etc.), weather and climate data, soil data, remote sensing data to determine the behavior and development of crops under different conditions. Applications differ considerably depending on the focus of the involved researchers and practitioners. However, some specific characteristics are particularly crucial for the work of this community. First, data used is generally highly heterogeneous, coming from a variety of sources and implementing different standards. Moreover, it is common that at least part of it concerns spatiotemporal information and thus specific analytics tools are required that can handle such data. For larger scale applications, for instance for assessments of larger geographical regions on high spatiotemporal resolutions that need to be finalized within time requirements, commonly used computing environments (such as laptops and single desktop computers or isolated modelling servers) might deliver insufficient resources. In such cases, parallel and/or distributed computing facilities can be considered to improve overall computing power.

To be able to cope better with such new challenges, the agro-climatic modelling use case in AGINFRA+ focuses on harnessing large scale modelling exercises, requiring substantial computing resources. The network of distributed computing resources is used to run such models in a performant manner, using distributed and parallel computing techniques. Besides, it exploits the typical open science related characteristics of the D4Science environment. Collaborative modelling, where teams of researchers work together to first develop and test models in literate programming environments and then deploy and openly share the developed algorithms can be an important step towards open science. In that way, complying with FAIR principles exceeds the level of only data sharing and reuse and adds the FAIR publication of data science algorithms.

3.2 Use case – crop growth modelling

One of the AGINFRA+ use cases that was explored and for which typical research applications were implemented and deployed into a VRE, with the aim to demonstrate and evaluate the usability for users of the agro-climatic modelling community, is crop growth modelling. Simulations using crop growth models are one of the important components in yield forecasting, used frequently in food security research and related research areas. Currently, the application of crop growth models is often still limited by the available computing resources. The application piloted in AGINFRA+, applying

European or global scale crop simulations on the detailed level of agricultural parcels is currently too demanding for many existing research infrastructures in the field. To meet the requirements for such large scale, high-resolution crop growth modelling exercises, the following facilities are indispensable: efficient retrieval of spatiotemporal data streams, spatiotemporal data wrangling and data processing, running models at scale using distributed computing resources and parallel technologies computing, and intuitive spatiotemporal visualization.

In the AGINFRA+ crop growth modelling use case, the preprocessing of spatiotemporal data is an integral part of the AgroDataCube infrastructure. This infrastructure provides Dutch agricultural open data as a service to research and business. The AgroDataCube ingests and merges different spatiotemporal data streams that are relevant for agricultural and environmental applications (among others weather data, agronomic data, parcel geometries, Sentinel-2 satellite data, and soil data). It provides a set of well-documented, ready to use REST services that allow retrieval of the merged data on the parcel level in usable packages and a standardized format (GeoJSON). To cope with the requirement to scale up simulations, the widely used WOFOST crop growth simulation model [18] was embedded into a distributed computing framework that facilitates the distribution of computing jobs over a compute cluster. The resulting modules have been integrated into the VRE as DataMiner algorithms [19], and were published, using the D4Science catalogue service, to make them discoverable and reusable as FAIR algorithms for the whole community. As the requirements for spatiotemporal visualization in this use case were high, a dedicated visualization dashboard, developed on the basis of various VRE components, was developed (see Figure 2).

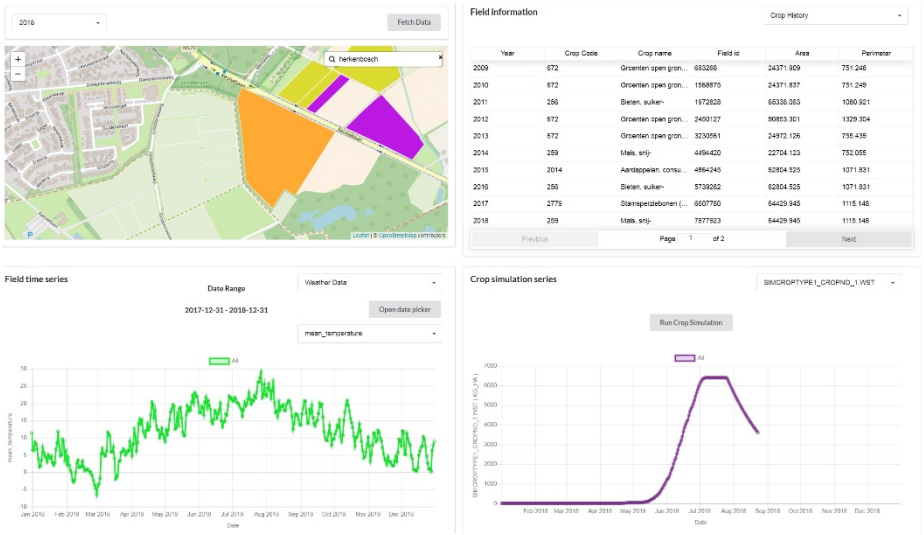

Fig. 2. AGINFRA+ crop growth modelling dashboard

For visual inspections, the dashboard allows geo-spatial and temporal visualization of the various data services provided through the AgroDataCube services that are input

to the crop growth simulations. Moreover, it offers its users the opportunity to manually search for, and select a specific agricultural parcel and initiate a WOFOST crop simulation by executing the VRE DataMiner algorithm using input data based on the selected field.

Generated simulation results are stored on the VRE's shared workspace, and the simulated parameters can be visualized as graphs and can be compared and analyzed side by side with the used input data. After being tested and quality checked, the developed models and algorithms and the generated output data can be shared with the broader user community, by publishing them through the VRE's catalogue service. Thus, the VRE is complying with the requirements of FAIR data services and open science in general, adding to that the opportunity to also share algorithms and models in a FAIR manner.

4 Conclusions and recommendations

AGINFRA+ has explored the usability of the VRE core services and building blocks offered through the D4Science platform for different scientific communities in the agri-food domain, by implementing a range of open science use cases that are typical for the agri-food research community. To demonstrate its value for the agro-climatic modelling community, AGINFRA+ has developed and deployed a pilot application around crop growth modelling, showing how large scale, high resolution crop growth modelling can be implemented by means of a VRE and its underlying core features for collaboration, computing, visualization and publication. The pilot successfully demonstrates how the D4Science infrastructure and the services deployed in the AGINFRA+ VRE can be used by distributed research teams to co-develop modelling workflows. It also shows that the offered cloud based computing and storage technologies are suited to efficiently scale up crop growth modelling, for larger geographic regions and on high-resolution parcel level. While the pilot demonstrates this specifically for the WOFOST model, the developed demonstrator can be used as a template to achieve similar results with many models in the agri-environmental and other domains.

Currently AGINFRA+ is trialing its use cases with a group of potential end users from its communities, allowing them to work with the developed pilots and to test and evaluate them on various criteria: ease of use, usefulness, openness, FAIRness and learning curve. The results of these evaluations will be used to further improve the developed tools and the underlying VRE services.

Acknowledgment
This work has received funding from the European Union's Horizon 2020 research and innovation programme under the AGINFRA PLUS project (grant agreement No 731001).

5 References.

1. United Nations, Sustainable Development Goals. https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals. Accessed 2019.09.29
2. European Commission, H2020 Societal Challenges. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges. Accessed 2019.09.29
3. Lokers R, Knapen R, Janssen S, van Randen Y, Jansen J (2016) Analysis of Big Data technologies for use in agro-environmental science. Environmental Modelling & Software 84:494-504. doi:10.1016/j.envsoft.2016.07.017
4. Jones S, Abramatic JF, (Eds) (2019) European Open Science Cloud (EOSC) Strategic Implementation Plan.
5. Barker M, Olabarriaga S, Wilkins-Diehr N, Gesing S, Katz DS, Shahand S, Henwood S, Glatard T, Jeffery K, Corrie B, Treloar A, Glaves H, Wyborn LAI, Chue Hong N, Costa A (2019) The global impact of science gateways, virtual research environments and virtual laboratories. Future Generation Computer Systems 95. doi:10.1016/j.future.2018.12.026
6. Ahmed I, Poole M, Trudeau A (2018) A Typology of Virtual Research Environments. doi:10.24251/hicss.2018.087
7. Zuiderwijk A (2017) Analysing Open Data in Virtual Research Environments: New Collaboration Opportunities to Improve Policy Making. International Journal of Electronic Government Research (IJEGR) 13 (4):76-92. doi:10.4018/ijegr.2017100105
8. Zuiderwijk A, Jeffery K, Bailo D, Yin Y (2016) Using Open Research Data for Public Policy Making: Opportunities of Virtual Research Environments. doi:10.1109/CeDEM.2016.20
9. Maier MW (1996) Architecting Principles for Systems-of-Systems. INCOSE International Symposium 6 (1):565-573. doi:10.1002/j.2334-5837.1996.tb02054.x
10. D4Science Consortium. D4Science: an e-infrastructure supporting virtual re-search environments". www.d4science.org.
11. EGI Foundation. "EGI e-infrastructure". www.egi.eu.
12. OpenAIRE Consortium. "OpenAIRE: the european scholarly communication data infrastructure". www.openaire.eu.
13. Janssen H, Janssen SJC, Knapen MJR, Meijninger WML, van Randen Y, la Riviere IJ, Roerink GJ (2018) AgroDataCube: A Big Open Data collection for Agri-Food Applications. Wageningen Environmental Research. doi:10.18174/455759
14. Caracciolo C, Stellato A, Morshed A, Johannsen G, Rajbhandari S, Jaques Y, Keizer J (2013) The AGROVOC linked dataset. Semantic Web Journal. doi:10.3233/sw-130106
15. German Federal Institute for Risk Assessment. "Foodrisk-labs". https://foodrisklabs.bfr.bund.de/foodrisk-labs/.
16. Assante M, Candela L, Castelli D, Cirillo R, Coro G, Frosini L, Lelii L, Mangiacrapa F, Pagano P, Panichi G, Sinibaldi F (2019) Enacting open science by D4Science. Future Generation Computer Systems. doi:10.1016/j.future.2019.05.063
17. Assante M, Boizet A, Candela L, Castelli D, Cirillo R, Coro G, Fernandez E, Filter M, Frosini L, Kakaletris G, Katsivelis P, Knapen MJR, Lelii L, Lokers RM, Mangiacrapa F, Pagano P, Panichi G, Penev L, Sinibaldi F, Zervas P (2019) Realising a Science Gateway for the Agri-food:

the AGINFRA PLUS Experience. 11th International Workshop on Science Gateways (IWSG 2019)
18. de Wit A, Boogaard H, Fumagalli D, Janssen S, Knapen R, van Kraalingen D, Supit I, van der Wijngaart R, van Diepen K (2019) 25 years of the WOFOST cropping systems model. Agricultural Systems 168:154-167. doi:10.1016/j.agsy.2018.06.018
19. Coro G, Panichi G, Scarponi P, Pagano P (2017) Cloud computing in a distributed e-infrastructure using the web processing service standard. Concurrency and Computation: Practice and Experience:e4219. doi:10.1002/cpe.4219