A comparative analysis of ORganic and Conventional Agriculture's impact on aquatic biodiversity (ORCA; BR/175/A1/ORCA) Data Management Plan – Initial version

De Wever, A.[1], Cours, M.[1], Milotic, T.[2], Lemmens, P.[3], Adriaens, D.[2], Denys, L.[2], Martens, K.[1], De Meester, L.[3]

[1] Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium – Aaike De Wever, orcid.org/0000-0001-8804-8321; Marie Cours; Koen Martens, orcid.org/0000-0001-8680-973X
[2] Research Institute for Nature and Forest, Brussels, Belgium – Tanja Milotic, orcid.org/0000-0002-3129-619; Dries Adriaens, orcid.org/0000-0002-7216-3111; Luc Denys
[3] Ecology Evolution Biodiversity Conservation, KULeuven, Leuven, Belgium – Pieter Lemmens, orcid.org/0000-0002-3135-9724; Luc De Meester, orcid.org/0000-0001-5433-6843

Plan details

Grant number: BR/175/A1/ORCA
Project Name: A comparative analysis of ORganic and Conventional Agriculture's impact on aquatic biodiversity (ORCA)
Project website: bio.kuleuven.be/eeb/ldm/ORCA

Principal Investigator:
Luc De Meester, KULeuven, Ecology Evolution Biodiversity Conservation, Leuven, Belgium.

Plan data contacts:
Aaike De Wever <aaike.dewever@naturalsciences.be>
Marie Cours <mcours@naturalsciences.be>
Tanja Milotic <tanja.milotic@inbo.be>

Description: Data Management Plan of the Belspo BRAIN project "A comparative analysis of ORganic and Conventional Agriculture's impact on aquatic biodiversity" (ORCA). This project investigates the combined effects of agriculture type (organic compared to conventional) and land use intensity (extensive versus intensive) on aquatic biodiversity in ponds and shallow lakes. The data underlying the planned analyses will be obtained from stratified sampling of farmland ponds, a common garden experiment, and through access to existing datasets.

Funder: Belgian Science Policy Office (Belspo)
Status of plan: Initial version, January 2019

Data set reference and name

ORCA freshwater database

The title “ORCA freshwater database” is a working title for the database that will constructed in the course of the “A comparative analysis of ORganic and Conventional Agriculture’s impact on aquatic biodiversity” (ORCA) project. We may choose to modify it or add a subtitle at a later stage. In addition to this “overall database”, data will likely be organised into (sub)datasets according to different types of analyses (e.g. “ORCA-Chemistry” and “ORCA-Macroinvertebrates”). This data management plan aims to cover both these (sub)datasets as well as the overall database. Note that this plan was compiled after the sampling campaign was largely completed, so topics such as field protocols and sample labeling in the field were already fixed at an earlier stage and will not be covered in detail in this document.

Planned update cycle

The current document is the initial version of the ORCA Data Management Plan. A revision is planned for autumn 2019. By then data compilation will be in an advanced stage and it should be feasible to plan publication and public release of the generated data in more detail.

Data set description

Project background

Agriculture is one of the most pervasive human activities on earth, impacting key natural resources (Rockström et al. 2009, Steffen et al. 2015), e.g. through eutrophication and the use of pesticides (Smith et al. 1999, Beketov et al. 2013). The impact of organic farming is expected to differ from that of conventional farming because only organic fertilizers are used, the strongly restricted use of a limited set of pesticides, and the protection of more natural elements such as ponds from adverse effects by larger buffer zones. In the ORCA project we use pond systems to investigate and assess the impact of agricultural on aquatic biodiversity in a comparative analysis of organic and conventional agricultural practices.

For this purpose, we will combine multiple existing databases with newly collected data;

(1) Sampling localities for collecting additional data were selected to ensure sufficient coverage of each type of agriculture (organic versus conventional) and a range in land-use intensity (extensive versus intensive) to enable comparative analysis of their combined effects on aquatic biodiversity in ponds. More specifically, we have sampled ponds along a gradient of cropland versus grassland-dominated areas embedded (to various degree) in organic (26 ponds) or conventional farmland (22 ponds; i.e. a total of 48 sampled ponds). We collected data on a broad range of environmental variables (>20 variables, cf. De Bie et al. 2012), amphibians, fish, macrophytes, macro-invertebrates, and zooplankton.
(2) In addition, we use existing data collected in previous projects using similar sampling techniques, in particular those compiled during the Belspo BRAIN project SAFRED (Saving Freshwater Biodiversity Research Data).

Using the newly collected data —following the stratified sampling design outlined above— in combination with existing data, we will assess the effects of different types of agricultural practices

on taxonomic, functional and intraspecific genetic diversity of multiple aquatic organism groups at local and regional spatial scales.

Newly generated data

We performed detailed environmental characterisation of the sites selected for sampling and quantify community composition and diversity of five different aquatic organism groups: zooplankton, macro-invertebrates, macrophytes, amphibians, and fish. These groups are both key targets for biodiversity conservation and important determinants of the ecological functioning and integrity of standing waters. For instance, macrophytes, zooplankton and fish are pivotal players in the context of regime shifts among alternative stable states in ponds and shallow lakes (Scheffer et al. 1993, Scheffer et al. 2001). Macro-invertebrates, zooplankton and macrophytes will be sampled and quantified using well-established methods. Fish and amphibian community composition and diversity will be analysed using environmental DNA (eDNA) metabarcoding as this provides a more reliable and less intrusive way to screen for their presence and relative abundance.

In summary, the following data will be generated:

- General overview of selected ponds and sampling events
- Field observations on pond characteristics such as surrounding land use and management
- Field measurements of pH, EC, temperature and oxygen
- Photographic documentation of ponds and sampling sites
- Landscape and land use characteristics of selected ponds, as derived from existing GIS layers
- Overview of collected samples
- Laboratory measurements of dry and suspended matter (KULeuven)
- Laboratory measurements of major ions and nutrients (RBINS)
- Zooplankton community composition and diversity – counts
- Macro-invertebrate community composition and diversity – counts
- Macrophytes community composition and diversity – field survey, cover data
- Fish and amphibia community composition and diversity – eDNA analyses
- Data related to the common garden experiment
- Population genomic and quantitative genetic data

Details on each of these "data types" including the variables measured and responsible partners/persons are provided in Annex 1: Overview of different data types.

Existing data

The key existing data on which the project will build fall into two categories. On the one hand there is a wide range of GIS layers such as the Biological Valuation Map, yearly crop registration, rural development program/agri-environment measures as well as the recently created reference GIS-layer – the so called *'blue layer'* – with all waterbodies >10 m^2 (see Watervlakken versie 1.0[1]; Packet et al 2018). On the other hand, data from earlier pond projects compiled in the course of the Belspo BRAIN project SAFRED (Saving Freshwater Biodiversity Research Data) will be used in the integrated data analysis.

[1] http://www.geopunt.be/catalogus/datasetfolder/10e87ad3-8235-40e0-8269-42c3c96a884d https://metadata.geopunt.be/zoekdienst/apps/tabsearch/?uuid=10e87ad3-8235-40e0-8269-42c3c96a884d.

These GIS layers were used to identify areas dominated by organic versus conventional farming and stratify them according to land use type (land use dominated by grassland versus cropland) for selecting the pools for sampling. Based on existing GIS data-layers, we can also extract information on e.g. the land use patterns around the water bodies selected for sampling as well as those studied in previous projects (in the SAFRED database) covered by the ORCA freshwater database, but feasibility remains to be evaluated. Information that can be extracted from GIS layers includes agricultural practice, land use type (grassland versus cropland) at multiple spatial scales (from 50 to 3200 m), former land use (e.g. crop type information of last six years; year since conversion to organic agriculture) and the extent of buffer strips that may separate the pond from cultivated parcels (but again, feasibility to be evaluated). The latter will allow to evaluate the impact of buffer strips on pond quality and biodiversity, and this for ponds in areas dominated by organic and by conventional farmland. In addition to the data extraction from GIS layers, we also envisage to use the identifiers of Watervlakken 1.0 for identifying the water bodies and linking data from different analyses.

The data from previous projects including the BELSPO projects Manscape, Pondscape and Tommelen, the EU funded project BIOMAN (EVK-CT-1999-00046), the *Toegepast Wetenschappelijk Onderzoek Leefmilieu (TWOL)* project Midden-Limburg, National Fund of Scientific Research Flanders project De Maten (grant G.0358.01), and the BiodivERsA project Tippingpond were compiled and standardised (to the Darwin Core standard; tdwg.org/standards/dwc) in the course of the SAFRED project. These datasets were made available through data.freshwaterbiodiversity.eu/ipt and the Global Biodiversity Information Facility (GBIF; gbif.org) network, thus making them publicly available for use in overarching analyses of biodiversity patterns.
During the course of 2018 – early 2019, data on pond ecology and biodiversity of IAP project SPEEDY will also be incorporated into the SAFRED/Freshwater Information Platform database. This combined dataset already holds pond biodiversity data on multiple taxonomic groups along gradients of land use intensity from a large number of ponds (n > 350) (e.g. Declerck et al. 2006; De Bie et al. 2012), which offers rich possibilities to test hypotheses on the impact of land-use intensity and agricultural practices on biodiversity and its functional consequences (cf. functional diversity).

In summary, key existing data consists of

- GIS layers
    - 'Watervlakken 1.0' of lentic waterbodies in Flanders (as small as 1.5 m^2)
    - Biological Valuation Map (Dataset 2016, also available on Geopunt)
    - Yearly crop registration (Data available from AGIV, 2008-2017)
    - Rural development program/agri-environment measures
- Existing pond data from previous projects
    - already included in the SAFRED database: Manscape, Pondscape, BIOMAN (Belgium), Midden-Limburg, Tommelen, De Maten
    - data for future inclusion in the SAFRED database: SPEEDY

Centralising and integrating new and existing data

To ensure the efficient centralisation and integration of newly generated data and existing data we aim to discuss the data flows from the different labs and analysis types. Based on this we aim to work out a workflow to ensure that information collected in the field, during the analyses, from past projects and extracted from GIS layers gets integrated in an overall database which is made available for all partners and enables them to perform integrated analyses and produce products such as the "priority map on areas where biodiversity gains can be achieved by land use change" (Work package 4).

Topics to be considered as part of this workflow are discussed below, along with the agreement reached during the partner meeting of October 4th 2018:

- Responsibilities and timeline for delivery of different types of "raw data" (see also Appendix 1) and data extracted from the GIS layers.
    - Initiated, will be followed-up by RBINS (Marie Cours)
- Required minimal information (metadata) to be delivered along with the data files (see also "Standards and metadata")
    - See next section.
- Choice of a central repository for data exchange among partners. This can be cloud storage such as Google Drive/Dropbox/Onedrive.
    - Likely OneDrive @ RBINS, to be established.
- Agreement on naming conventions, file formats and organisation of files. Recommendations on these topics can be found in the Data Management guidelines worked out during SAFRED (tinyurl.com/y7olom3e).
    - Not yet discussed in detail, to be established.
- Data quality checks to be performed before data integration.
    - Not yet discussed in detail, to be established.
- Timeline for integrating the compiled data in an "overall ORCA freshwater database".
    - Not yet discussed in detail, to be established.
- Producing export queries for publishing biodiversity data in Darwin Core format (see also "Data sharing").
    - Agreed, will be taken up by RBINS, but will require support from BBPf/INBO.

Background on the ORCA freshwater database

Different (sub)datasets are currently (Autumn 2018) being compiled. Both the structure/tables and technology Access/SQLite/PostgreSQL for the ORCA freshwater database remain to be discussed and decided. The aim is to report these choices, along with summary statistics (number of sites and samples) here in a next iteration of the Data Management Plan.

Standards and metadata

Information on the datasets (metadata)

As outlined in the project proposal, background information on sampling localities and protocols will be recorded along with other metadata, and kept together with the actual data. In practice, the information on the (sub)datasets will initially be captured in ReadMe-files[2] covering the topics outlined below and will be compiled and provided along with the raw data for integration in the overall ORCA freshwater database. Metadata of the overall ORCA freshwater database or selected components considered for (separate) publication will be entered or transferred to the Freshwater Metadatabase at data.freshwaterbiodiversity.eu/metadb (which is part of the SAFRED infrastructure), which supports metadata export in the Ecological Metadata Language (EML)

[2] A separate text file in .txt, .md or .rft format is preferred as these file types are universally readable and don't require software licenses. The .md format is also readable by general purpose data repositories such as Zenodo (you only need to upload the .md file and the content is immediately shown as documentation for the uploaded data). A ReadMe-tab in an Excel-file would also be accepted for data originally recorded in Excel, but in this case metadata needs to be be extracted into a separate file during further processing.

standard. For existing datasets (i.e. those in the scope for the SAFRED project), metadata for component datasets will be imported/integrated from existing sources where available. If this is not the case, metadata will be encoded by project partners.

Minimal information on dataset (metadata) required includes:

- Title/Name of the dataset
- Corresponding filename of the dataset
- Short description
- Keywords
- Contact person(s)
- Description of fields/column names including measurement units
- Sampling methods/protocol or reference of paper containing method description
- Time frame: sampling period
- Version

Use of standard control vocabularies

At this stage (August 2018), the use of control vocabulary (standard list of terms) has not yet been considered. Its relevance for e.g. pond characterisation (e.g. pond morphology, sediment characteristics, etc.) will be considered. The unique identifiers of Watervlakken 1.0 provide a standard reference for identifying the water bodies and linking data from different analyses.

Standards for organising and sharing data

The database structure of the ORCA freshwater database will be reviewed in the course of the project, considering the need for creating export queries matching the Darwin Core standard (tdwg.org/standards/dwc) at a later stage. In parallel, we will identify the Darwin Core terms/fields (rs.tdwg.org/dwc/terms) which are most relevant in the context of this database.

Biodiversity data that will be released in the framework of ORCA will be exported in Darwin Core format and mapped through GBIF's Integrated Publishing Toolkit (IPT) software (see details under data sharing).

Data sharing

As explained under "Existing data", the data from former projects as compiled during the SAFRED project are already publicly available. Newly generated data will be made available to project partners in the framework of joint analyses as outlined in the section "Centralising and integrating new and existing data". The construction of this Data Management Plan is considered as a tool to discuss and streamline the integration of the newly generated data and facilitate its integration in the SAFRED infrastructure at RBINS and public release of the data (after an embargo period allowing for processing and scientific publication based on the data). Note however that the data publication can be prepared well in advance for release at a later stage and project partners have the obligation to make the data publicly available at the latest in parallel to the associated scientific papers. Preparatory work for publication of data should be done by the end of the project. Data publication will focus on biodiversity related data using the tools available through the Global Biodiversity Information Facility (gbif.org) network, thereby referring to the workflow worked out during SAFRED and adopting the supported options for data citation and licensing.

Details on a standard embargo period and the procedure for release of the newly generated remain to be discussed taking the planned data analyses and related scientific papers into consideration. In addition to the data integration in the SAFRED infrastructure (and hence the GBIF network), we will consider data archiving through a general-purpose repository such as Dryad (datadryad.org) for the overall dataset in parallel to its release as a data paper.

Detailed tasks and responsibilities with regards to data sharing/publication remain to be discussed and documented.

Archiving and preservation (including storage and backup)

Backup of (sub)datasets

During the compilation of the raw data, it is the project partners' responsibility to ensure that the data are properly backed up. As explained in the Data Management guidelines worked out during SAFRED (tinyurl.com/y7olom3e – see Data security section), we recommend applying the 3-2-1 rule: 3 copies, 2 media, 1 offsite copy. As data files for most types of data are typically >10Mb (with the possible exception of the eDNA analysis results), file size and storage capacity are unlikely to pose any problems.

Archiving of overall ORCA freshwater database

Organising the back-up and archival of the centralised files and overall ORCA freshwater database is the responsibility of RBINS. Options for long-term archival at an institutional archive and/or a general purpose repository (such as Dryad–datadryad.org for data associated with scientific papers) or Zenodo (zenodo.org) remain to be discussed among the partners, as is the deposition of eDNA data in one of the INSDC (International Nucleotide Sequence Database Collaboration) databases (such as the EMBL-EBI or NCBI-GenBank).

Archiving of ORCA freshwater samples

Physical macroinvertebrate and zooplankton samples collected during the ORCA project will be incorporated in the collections of the Royal Belgian Institute of Natural Sciences (RBINS). To this end, the inventory number (Coll. RBINS IG 33.632) has been assigned to this collection and a label with this number will be attached to the ORCA samples during processing. Once analysis of the samples in the framework of ORCA has been finalised, they can be transferred to the collection for long-term archival.

Topics to consider for next iteration of the ORCA Data Management Plan

Several potentially relevant issues are not yet addressed in this initial version of the Data Management Plan, these could be considered during next iteration. These include:

- Ethics and legal compliance – Are there any ethical or legal issues that can have an impact on data sharing? E.g. how to deal with data collected on the property of private land owners, and permissions obtained for data collection.
- Applicable data policy/institutional policies such as Belspo regulations for projects and institutional policies such as the INBO Open Data policy
- Data sharing among researchers during the project and ensuring that everyone is using the latest version
- Versioning, exchanging and archival of analysis scripts during the project? (e.g. using GitHub)
- Details on “Population genomic and quantitative genetic analyses” data

References

Beketov, M. A., B. J. Kefford, R. B. Schäfer, and M. Liess. 2013. Pesticides reduce regional biodiversity of stream invertebrates. Proceedings of the National Academy of Sciences USA 110:11039-11043.

De Bie, T., L. De Meester, L. Brendonck, K. Martens, B. Goddeeris, D. Ercken, H. Hampel, L. Denys, L. Vanhecke, K. Van der Gucht, J. Van Wichelen, W. Vyverman, and S. A. J. Declerck. 2012. Body size and dispersal mode as key traits determining metacommunity structure of aquatic organisms. Ecology Letters 15:740-747.

Declerck, S., T. De Bie, D. Ercken, H. Hampel, S. Schrijvers, J. Van Wichelen, V. Gillard, R. Mandiki, B. Losson, D. Bauwens, S. Keijers, W. Vyverman, B. Goddeeris, L. De Meester, L. Brendonck, and K. Martens. 2006. Ecological characteristics of small farmland ponds: Associations with land use practices at multiple spatial scales. Biological Conservation 131:523-532.

Packet J., Scheers K., Smeekens V., Leyssen A., Wils C., Denys L. 2018. Watervlakken versie 1.0: polygonenkaart van stilstaand water in Vlaanderen. Een nieuw instrument voor onderzoek, water-, milieu- en natuurbeleid. Rapporten Instituut voor Natuur- en Bosonderzoek INBO.R.2018.14, 43 p.

Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, and H. J. Schellnhuber. 2009. A safe operating space for humanity. Nature 461:472-475.

Scheffer, M., S. Carpenter, J. A. Foley, C. Folke, and B. Walker. 2001. Catastrophic shifts in ecosystems. Nature 413:591-596.

Scheffer, M., S. H. Hosper, M. L. Meijer, B. Moss, and E. Jeppesen. 1993. Alternative equilibria in shallow lakes. Trends in Ecology & Evolution 8:275-279.

Smith, V. H., G. D. Tilman, and J. C. Nekola. 1999. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental Pollution 100:179-196.

Steffen, W., K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, and C. A. de Wit. 2015. Planetary boundaries: Guiding human development on a changing planet. Science 347:1259855.

Annex 1: Overview of different data types

Data types: description of available/expected data & responsible parties

Description	Type	File format	File/Folder name	Responsible partner	Responsible/Contact person	Planned availability for partners
General overview of selected ponds and sampling events	Sampling event details	Excel	sampled_ponds_ORCA.xlsx	RBINS	Marie Cours	
Field observations on pond characteristics such as surrounding land use and management	Sampling event details	PDF	ORCA_RAW DATA_20171005 (1).xlsx	RBINS	Marie Cours	March 2018
Field measurements of pH, EC, temperature and oxygen	Measurements	Excel	ORCA_FieldMeasurements	KU Leuven	Rafaela Almeida	March 2018
Photographic documentation of ponds and sampling sites - images from 2017	Sampling event details		ORCA_Images2017	KBIN	Marie Cours	
Photographic documentation of ponds and sampling sites - images from 2018	Sampling event details		ORCA_Images2018	INBO	Rein Brys	October 2018
Landscape and land use characteristics of selected ponds, as derived from existing GIS layers	Measurements	Excel	ORCA_poelen_2017_landUse.xlsx	INBO	Dries Adriaens	May 2018
Data on presence and characteristics of bufferstrips (field + GIS)	Measurements / Sampling event details	Excel	ORCA_BufferStrips	INBO	Jo Packet	January 2019
Overview of collected samples	Sampling event details	Excel	ORCA_Samples	RBINS	Marie Cours	
Laboratory measurements of dry and suspended matter (KULeuven)	Measurements		ORCA_SUSPS		Rafaela Almeida	March 2018
Laboratory measurements of major ions and nutrients (RBINS)	Measurements	Excel	ORCA_WaterChemistry	RBINS	Marie Cours	
Zooplankton community composition and diversity – counts	Occurrences	Excel	ORCA_Zooplankton	KU Leuven	Rafaela Almeida	September 2018
Macro-invertebrate community composition and diversity – counts	Occurrences	Excel	ORCA_MacroInvertebrates	RBINS	Marie Cours	
Macrophytes community composition and diversity – field survey, cover data	Occurrences	Excel	ORCA_Macrophytes	INBO	Luc Denys/Jo Packet	March 2018
Fish community composition and diversity – eDNA analyses	Occurrences		ORCA_Fish	INBO	Rein Brys	February 2019
Amphibia community composition and diversity – eDNA analyses	Occurrences		ORCA_Amphibians	INBO	Rein Brys	February 2019
Data related to the common garden experiment			ORCA_CommonGarden	KU Leuven	Rafaela Almeida	May-June 2019
Population genomic and quantitative genetic data	Occurrences	FASTQ	ORCA_GEN	KU Leuven	Rafaela Almeida	End of 2019