Targeted Brainstorming

Powering a Unified Knowledge Base for Plant Natural Product Discovery

📅 **When**: 14 October, 14:00 – 16:00

📍 **Where**: Radix Building, Wageningen Campus (B:107 R:M1.160 Meetingroom M2, Radix Mid), Droevendaalsesteeg 1, Building 107, 6708 PB Wageningen

Participants

Symposium Speakers

1. Willighagen, Egon (TGX) <egon.willighagen@maastrichtuniversity.nl>
2. Slenter, Denise (TGX) <denise.slenter@maastrichtuniversity.nl>
3. Rutz, Adriano <rutz@imsb.biol.ethz.ch>
4. Allard, Pierre-Marie <pierre-marie.allard@unifr.ch>
5. Louis-Felix Nothias <louis-felix.nothias@univ-cotedazur.fr>

Wageningen University, Bioinformatics and Biosystematics groups:

6. Elena Del Pup : Del Pup, Elena <elena.delpup@wur.nl>
7. Marnix Medema : Medema, Marnix <marnix.medema@wur.nl>
8. Justin van der Hooft : Hooft, Justin van der <justin.vanderhooft@wur.nl>
9. Mitja Zdouc : Zdouc, Mitja <mitja.zdouc@wur.nl>

External invitees:

10. Tito Damiani : Tito Damiani <tito.damiani@uochb.cas.cz>

Introduction

Plant specialized metabolites are a major source of medicines, foods, and natural products. With the rapid growth of multi-omics and paired transcriptomics–metabolomics datasets, we now have unprecedented opportunities to uncover how plants produce these molecules. But today, insights from these datasets remain scattered across publications and tools, making it difficult to connect and reuse them.

Our vision is to build a harmonized Linked Open Data knowledge base that collects and connects experimentally validated pathways from various sources of evidence (transcriptomics, metabolomics, cheminformatics, pathway databases, and literature) and extends them with pathway predictions. Such a platform would allow us to:

- Query across datasets and evidence types to validate pathway predictions,
- Prioritize promising candidates for chemical and biosynthetic novelty,
- Continuously integrate new data contributed by the community,
- Provide an open, reusable foundation for both research and education.

Knowledge graphs and linked open data approaches give us a powerful way to discover plant pathways of natural products. They can integrate annotations from tools like plantMASST, MS2LDA, ENPKG, matchMS, and SIRIUS, connect them with pathway frameworks such as WikiPathways and PlantCyc, and enrich them with information from repositories like MIBiG, GNPS, Wikidata, and ChEMBL.

The goal of today's brainstorming is therefore to explore **how we can align existing resources, connect predictions to pathway knowledge, and design a participative platform** that enables researchers worldwide to query, validate, and build upon each other's pathway predictions.

By working together, we can lay the foundation for a community-driven infrastructure — one that not only collects predictions, but makes them explorable, reusable, and actionable for plant natural product discovery.

Goals of the brainstorming:

The aim of this session is to identify opportunities and challenges in building a unified, queryable knowledge base for plant natural product discovery. Specifically, we want to:

1. **Map existing resources and projects**
2. **Identify key challenges and bottlenecks**
3. **Discuss collaboration opportunities and define next steps toward a participative platform**

Pre-brainstorm Preparation

PLEASE ADD YOUR THOUGHTS BEFORE THE MEETING :)

1. What existing **resources and tools (including projects, databases, and standards)** are there to support automated discovery of plant biosynthesis and connected natural products?

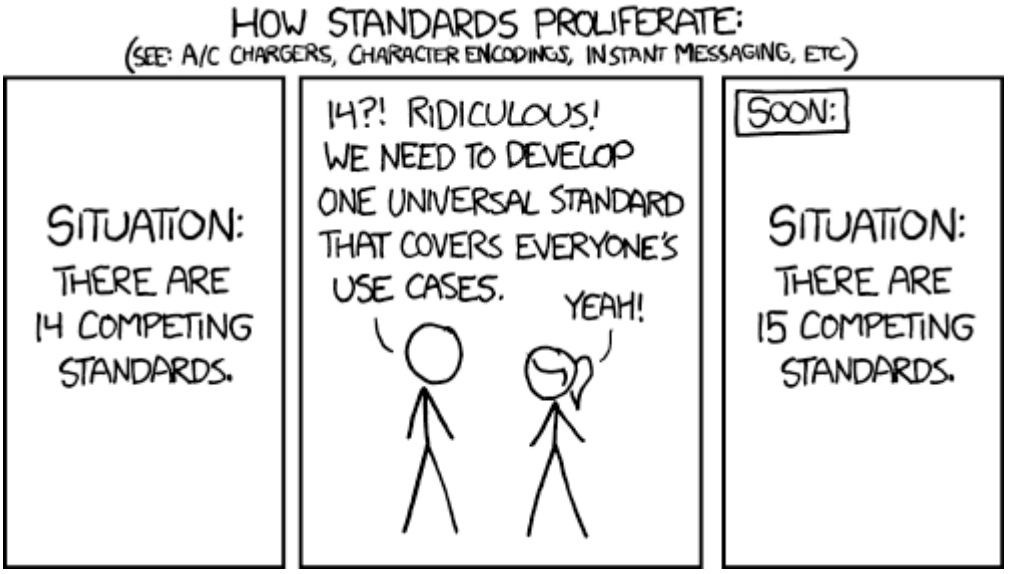
→ Please add to the collaborative: Linked omics Tools for plant biosynthesis

2. ***What do you see as the most challenging aspect of building a unified knowledge base for plant natural product discovery? :***
    1. *What do you see as the most critical **technical barrier** to connecting and querying plant pathway knowledge and predictions?*
    2. *What do you see as the most critical **biological challenge** in linking transcriptomics and metabolomics to pathways?*

3. *What do you see as the biggest **community or sustainability challenge** for a shared knowledge base?*

Please answer below before the meeting

Elena	**1. Technical challenge** How to integrate predictions: make predicted edges in a KG How to extend/update the KG collaboratively/iteratively **2. Biological challenge** Biological challenge: Tools developed for microorganisms often have to be specially adapted for plants: e.g. MIBIG is not suited to represent incomplete clustering in plants. Connecting metabolomics and transcriptomics is a challenge **3. Community-sustainability challenge** Develop minimum information standards for a plant pathway Coordinate with these different labs and maybe others? Apply for joint funding? Whitepaper? Interest in making something plant specific or better to include other kingdoms too?
Marnix	**1. Technical challenge** Faulty annotations of metabolite features. How to estimate likelihood of predictions to be correct. **2. Biological challenge** Timing of transcript/metabolite production is not always directly linked and metabolites accumulate. **3. Community-sustainability challenge** Long-term support/funding of curation (could be addressed through community-driven approaches).
Justin	**1. Technical challenge** Expression of reliability of information (predictions, inferred links, etc.,) Increasing amounts of data and metadata to work with **2. Biological challenge** Sparse data - how to work with missing data? Different degrees of clustering of genes. Different timings across omics layers (fast versus slow response) Missing proteomics layer Definition of pathways is a human construct (boundaries are not always clear) **3. Community-sustainability challenge** Financial support to maintain core functionality (who is paying?) Getting agreement on ontology/identifier/index to use
Tito	**1. Technical challenge** **2. Biological challenge** **3. Community-sustainability challenge**

	Long-term perpetual fundings. For example, I recently met Christoph Steinbeck who told me that the COCONUT database just obtained perpetual funding from the German government to be maintained.
Mitja	1. **Technical challenge** 2. **Biological challenge** 3. **Community-sustainability challenge** 
Adriano	1. **Technical challenge** None purely technical, rather interoperability due to the way humans structure(d) knowledge 2. **Biological challenge** Coverage 3. **Community-sustainability challenge** Strong advocacy to have a large community following. (also related to Tito/Mitja above: Securing funding to do useless things is pointless, coming with another failed standard without the community following it also)
Pierre-Marie	1. **Technical challenge** I don't see any technical challenges on the computational and data management side. Looks to me like we have most of the components of the toolbox or have ways to build them if they lack. Semantic web technologies are more than 30 years old, lightweight and not particularly complex to implement. Same regarding data acquisition. We now have ways to efficiently and cheaply profile large numbers of samples. We can't amplify metabolites !! (please find a solution for this so we can catch up with genomics :) 2. **Biological challenge** Metabolome is dynamic so this implies realizing numerous data acquisitions at different lifestages, in different environmental conditions and of different parts of the organism to achieve a comprehensive view of an organism. To be compared with large scale genomics characterization projects.

	3. Community-sustainability challenge Maybe it's here that are the biggest challenges. - How do we share common vocabularies ? Yes there are a lot of ontologies out there, but there will always be terms to be added, concepts to be refined etc. What are the best solutions and forums to discuss/propose these improvements and refinements. - How do we efficiently capture metadata ? - How to convince people that 5 stars in the LOD adoption scheme (https://5stardata.info/en/) is better than 4,3,2,1 or 0 ? Advantages to moving to more stars should be immediately perceived by researchers producing the data. How are researchers credited for moving their data to LOD ?
Denise	**1. Technical challenge** Technical challenges: annotations and interoperability between these **2. Biological challenge** Difference in size with respect to statistical analysis What information is needed for plant natural product discovery (e.g. metabolites, metabolic reactions, enzymes catalyzing these reactions, different species, genes, mappings between the latter two, and annotations for all of these). Provenance of data, and reproducibility of results. Evidence and weight of evidence. Model behind the knowledge graphs, pros and cons of different structures (think hyperedges). **3. Community-sustainability challenge** Maintenance; keeping data up-to-date Additional points from Denise: - automated/ queryable?
Egon	**1. Technical challenge** None. **2. Biological challenge** Measuring biology so that we can understand the processes in enough detail **3. Community-sustainability challenge** Lack of willingness to change the publishing system
Louis-Felix	**1. Technical challenge** **2. Biological challenge** **3. Community-sustainability challenge**

3. Would you be interested in participating in writing a white paper from the collective notes of the meeting? Preliminary title: "*How to build a knowledge base for plant natural product discovery*".
Please mention your name below as a potential contributor if that is the case:

- Elena Del Pup
- Marnix Medema
- Justin van der Hooft
- Mitja Zdouc
- Denise Slenter
- Pierre-Marie Allard
- Egon Willighagen
- please add your names to the list …

Brainstorming Questions (to be addressed during the session)

1. **Mapping the Landscape & Resources**

   - Given the map of the existing resources (Linked omics Tools for plant biosynthesis), how are these resources currently **connected** (if at all)? Are they interoperable? How to make them interoperable? What **key gaps remain/ what is missing**?

👉 Outcome: a **map of the ecosystem** (tools, datasets, standards).

---

2. **Challenges & Bottlenecks (biological, technical, community/sustainability)**

Look at the challenges identified before the meeting: group them and rank them for discussion.

- What are the hardest **biological challenges** in linking transcriptomics and metabolomics to validate pathway predictions?
- What **technical barriers** do we face (data formats, ontologies, standards, alignment, interoperability)?
- What **community or sustainability barriers** exist (adoption, curation, sustainability, incentives for sharing or for curating).

👉 *Outcome*: a **ranked challenge list** that can become the backbone of a white paper introduction / discussion and show that you're addressing real community-identified needs.

---

3. **Collaborations / Opportunities**

   - Where are synergies between our projects and expertise? Where do the participants' tools and expertise naturally complement each other? What concrete collaborations could we start to align predictions and build shared infrastructure?
   - What collaborations would have the biggest impact in the short term (6–12 months)? And in the long term?

- What concrete steps could we take now (pilot dataset, joint curation, grant writing)? What short-term outcomes could we target (e.g. white paper, joint pilot dataset, Plant WikiPathways community)?
- Would anyone be interested in hosting this meeting again in a year from now?
- Where will we collect and integrate semantically the diversity of research output?
-

👉 *Outcome*: This part produces **actionable next steps and names attached to them**, so the symposium is not just theoretical. It directly feeds into project and network building.

Potential outcomes - takeaways

2. **White paper** together on “How to build a knowledge base for plant natural product discovery”. Potential contributors:
   - Elena Del Pup
   - Marnix Medema
   - Justin van der Hooft
   - Mitja Zdouc
   - Pierre-Marie Allard
   - Adriano Rutz
   - Egon Willighagen
   - Denise Slenter
   - add your names to the list …

3. Final PhD chapter for Elena Del Pup : Linked Data Infrastructure for Plant Biosynthesis
   - pilot dataset?
   - discuss everyone’s involvement
4. Establishment of a Plant Wikipathway community

Agenda (afternoon)

Timekeeper:

12:30 – 13:50 Speakers’ lunch (Orion building, https://share.google/tIuC3Kr1nwCIRFDzz ~10 minute walk from Aurora)

13:50 – 14:00 Walk together to Radix building (~5 minutes walk)

14:00 – 14:10 *Introduction (Elena)*

- State the vision: connected, queryable knowledge base for plant pathways and integrating transcriptomics and metabolomics

- Explain the goals of the session (map resources, identify challenges, define collaborations) and the structure : where I am and where I want to go

    - make the current projects compatible with pathway information
    - power omics-based predictions/predicted pathways
    - useful takeaways from the morning symposium

14:10 – 14:30 *Mapping the Landscape (20 min max)*

- Output: shared map of tools/resources, connections, gaps

14:30 – 15:10 *Challenges & Bottlenecks (40 min max)*

- Output: ranked list of top challenges (technical, biological, social)

15:10 – 15:40 *Collaborations & Opportunities (30 min max)*

- Output: concrete collaboration ideas (pilot projects, white paper contributors, community launch steps)

15:40 – 16:00 *Wrap-up & Next Steps*

- Summarise key insights
- Agree on 1–2 immediate next steps (white paper draft, Plant WikiPathways community, pilot dataset)
- Assign follow-ups (Elena circulates notes & action items)

Note Keeping

Shared slides: EDP_Targeted Brainstorming (afternoon).pptx

Present: Pierre-Marie Allard, Adriano Rutz, Marnix Medema, Tito Damiani, Louis-Felix Nothias, Justin van der Hooft, Egon Willighagen, Denise Slender, Elena Del Pup, Mitja Zdouc

Overall idea: using Wikipathways for storing the information in the **GPML2021** ontology model

- Brief introduction by Elena:
    - Integrating Multiomics to annotate Plant Biosynthesis (PhD work)
    - Vision: community-driven knowledge base to prioritize NPs; objectives and challenges -> NP pathway predictions and ranking/prioritization
- Denise: What is actually a pathway? A series of reactions? Cluster of Metabolic conversions? Full network of metabolic reactions? Include protein-protein interactions?
- Tito: What are the objects that are connected?

- Tito/Elena: object of predictions are enzymes/genes, and metabolites (substrate, product)
- Elena: nodes can be genes, enzymes or other; using PathVisio
- Marnix: complex discussion what a pathway is, depending on context. Various possible definitions that are distinct: e.g., route from common precursors to final (exported) product, subnetwork within a larger network or unbranched linear module of enzymatic transformations. Clarity is important.
- Denise: Graphical Pathway Markup Language (GPML) is not the best model to collect all metabolic conversions in one model → splitting in separate models is better (for rendering the visualizations). Multi-species pathways come with the issue of "hairballs" (not visually appealing). Current model is GPML 2021, there is still some work to be done to make it work with the WikiPathways (website).
- Elena: wants to be able to predict genes that are currently missing in these networks, thus they need to be large to have sufficient data for AI prediction (?) → Address knowledge gaps.
    - ECO ontology has this: computational inference used in automatic assertion
- Egon/Denise: You can now add databases for identifier mappings without needing new releases of BridgeDB. Mappings of IDs are needed (needed for project: MIBiG?). Bioregistry and identifiers.org are integrated into BridgeDb to give IDs (URIs) their namespace. BridgeDb has an API that connects to existing databases and provides the ID mappings (also includes Wikidata).
- Mitja/Elena: Asked about the scope of Elena's envisioned project: wants to create a new resource that can create edges and include data that wikipathways cannot contain
- Denise: representation: sometimes the substrate does not have te full structure but spectral information/MS2… Sometimes the ChEBI is not known but it is just a theoretical compound.
- Egon: computationally derived conclusions. MetWare for doing that
- Marnix: what data resources do we have: enzyme reaction links. Substrate specificity and catalytic activity of the enzymes.
- Egon: Wikibase could be a platform to link for this.
- Mitja: MITE paper, what is the governance of the idea of the model and the resource. Who provides the service?
- Egon: Wikibase can be hosted by wikimedia foundation (DE)
- Adriano: PubChem they have RDF but does not expose it
- Egon: we host a mirror SPARQL endpoint for ChEMBL RDF data.
- PM: https://pmn.plantcyc.org/pathway?orgid=PLANT&id=PWYQT-4472&detail-level=4
- https://pmn.plantcyc.org/pathway?orgid=PLANT&id=PWY-5992
- PM: EC numbers connected to reaction rules (SMIRKS) and substrates (SMARTS) B-nice (Vassily Hatzimanikatis group, EPFL)?
- Denise: check for stereochemistry we could do systematically.
- ECO Evidence and Conclusion Ontology
- Chemical Evidence Ontology → https://github.com/semanticchemistry/semanticchemistry

Resources

- https://www.bgee.org/?action=access.
- MetWare
- MetaNetX https://www.metanetx.org/
- Oxo https://www.ebi.ac.uk/spot/oxo/
- LinkML https://linkml.io/

Challenges:

Technical challenges:

- Egon: tool to add predictions to GPML, and evidence and conclusion ontology (ECO) and reflect/encode uncertainty. SHACL and SHEX shapes. then make additional statements of that predicted edge.Chemical information ontology can be reused.
- Also distinguish between predictions external vs internal to the KG: predictions by external tools, or e.g. GNN-based predictions of links based on the KG itself.
- Adriano: predicted reactions can be part of wikidata? or wikipathways? But not for really specific predictions.
- Marnix: what is a prediction?
- Egon: If a prediction has been published it has support and is part of wikidata. Because also Wikipedia does not have support.
- Denise: it is like a nanopublication for theoretical data.
- Tito: wikidata does not accept only measurements like spectral data.
- PM: Idea to go to wikibase for this challenge, becasue of the experimental details.
- Marnix: ENPKG is starting from raw data and Wikidata is for experimental knowledge.
- Egon: CHEMBL is also not fully in Wikidata because of that; many theoretical compounds.
- Adriano: which edges are you trying to draw? Elena: new edges between genes.
- Key question: what is the minimal information to define (or predict) a pathway? Which modalities are involved? What additional data are most helpful in defining/predicting?
- Elena: Minimal information about a biosynthetic (plant) pathways.
- Adriano: do it with other organisms. (Is plant too niche?)
- PM: MIBiG: gene clusters (region on the chromosome).
- Marnix: complementary gene clusters (common in plants, fungi) that encode proteins for a specific pathway (set of reactions).
- Denise: what is the difference to gene ontology? See metabolic process Gene Ontology Term (GO:0008152)
- PM: ENPKG is a dataset.
- Justin: what is needed for a pathway, is not the same information as needed for predictions.
- Elena: distribution of GPMLs from PlantCyc is the next step.

Biological challenges:

- Marnix: transcriptomics, metabolomics: difficult for the dynamics and challenges.
- Denise: metabolome more dynamic than the transcriptome.
- Justin: sparsity of data/missing data + different dynamics across omics layers
- Tito: paired omics could be integrated?
- Tito: add also the predicted intermediate steps

- PM: identify end products
- Denise: big hairball of data and where substrate and products are all connected but sometimes also not really.
- Egon: the order of reactions is not known sometimes.
- Justin: definition of pathways is a human construct (boundaries are not always clear)
- Denise: data structure, add reaction node. Check is product from one pathway feeds into another pathway in PlantCyc (or does it contain disconnected reactions/edges).

- Adriano: Sparsity is in the heart of the model: not many known pathways
-
- PM**:** prune it at the end, and say.

- Metabolic Pathway versus Metabolic Reactions

- Can we define a pathway without (known) genomic information?

- Looking for prototype datasets

Sustainability challenges:

- Mitja: What is a suitable data structure? First predictions, then model?
- Whitepaper: make the topic more generic (not only for plants), but on natural product discovery (using MERMAID to visualize the minimal model).

- Biosynthetic reaction

Action points/takeaways:

- Minimum information: what is a pathway, what is a prediction?
- Governance

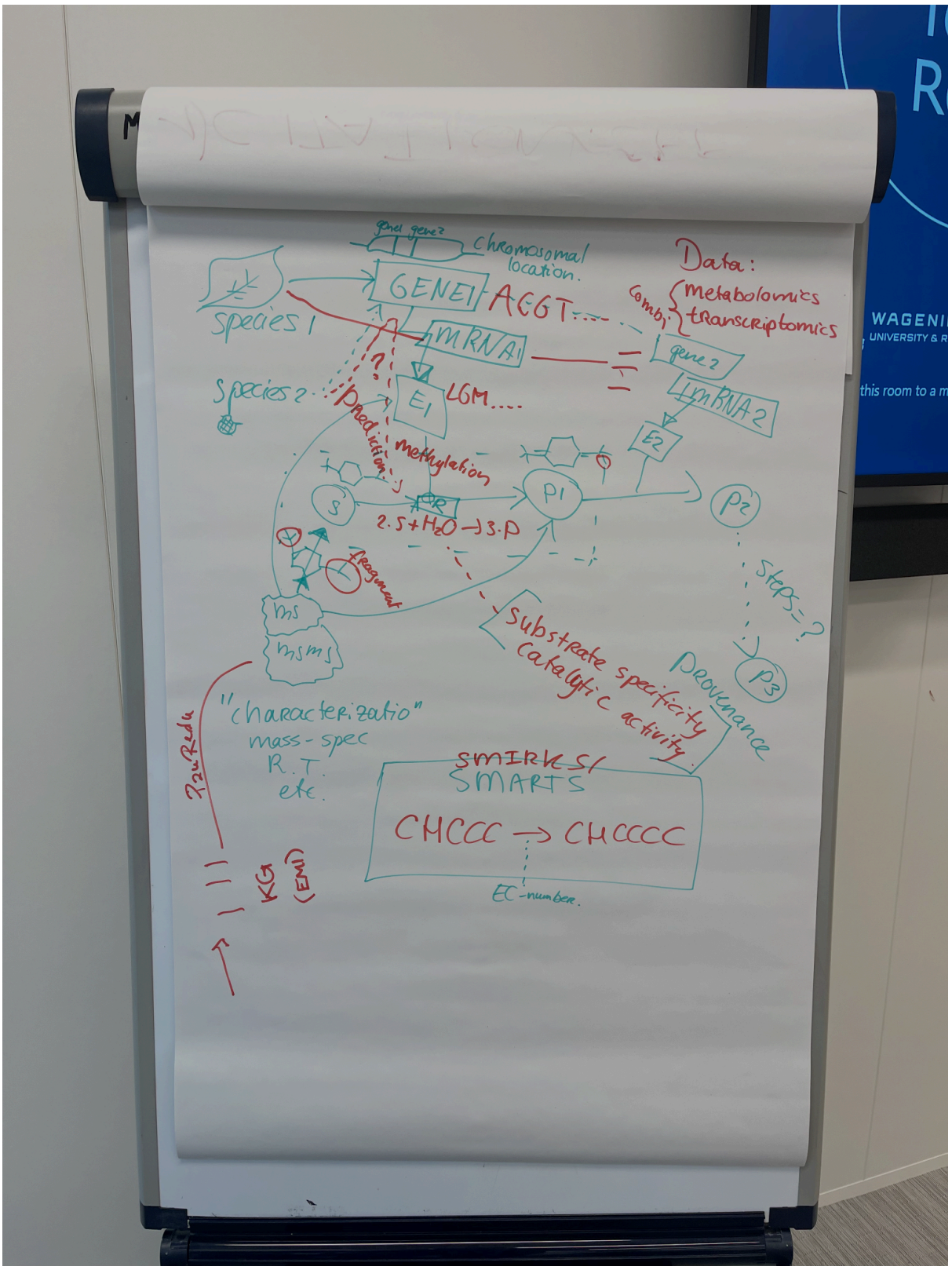
Involvement

- Denise: github actions and github pages. Involvement with industry.
- Egon: gpml to ttl file
- Mitja: MITE schema was taking much longer than expecting.
- Adriano: happy to help with whatever I can.
- PM: whitepaper, with repository where we brainstorm. LinkML to collaboratively edit ontologies and reusing and checking all the terms and identify places if we need to add things.
- Denise: OXO from EBI is a good way to link ontologies, cross ontologies instead of creating new ones. To see which ontology is covering most of the needed things, and what are the connections between the ontologies.
- Egon: LinkML is to describe the data format.
- Justin: setting up a community based page for long-term community building
- Egon/Justin: for within PhD work, work with what you currently have (and make/keep the framework extendable and flexible), i.e., some discussions (on definitions) can take a long time
- Marnix: yes to follow up meetings and make a plan for this.

Other people to involve

- Jean Luc Wolfender?
- Daniel Probst?
- Yasin?
- Kumar?

With regard to the white paper, I would suggest making an effort to structure the notes and use them to create a potential draft structure (as a basis for discussion), share this and then plan a joint video call with everyone


gene1 gene2
chromosomal location.
GENE1 ACGT....
species 1
species 2
mRNA1
E1
LGM....
Prediction
methylation
S
R
P1
P2
P3
2.5+H2O → 3.P
fragment
ms
msms
Data:
metabolomics
transcriptomics
Combi
gene2
LynRNA2
E2
steps=?
Provenance
Substrate specificity
Catalytic activity
SMIRKS/
SMARTS
CHCCC → CHCCCC
EC-number
"characterization"
mass-spec
R.T.
etc.
KG
(EMI)