

Semantyczne modelowanie danych doświadczalnych i jego zastosowanie w analizie wyników fenotypowania roślin



Zakład Biometrii i Bioinformatyki
Instytut Genetyki Roślin
Polskiej Akademii Nauk
w Poznaniu

mgr inż. Hanna Ćwiek-Kupczyńska

ORCID 0000-0001-9113-567X

hcwi@igr.poznan.pl

Promotor: prof. dr hab. Paweł Krajewski

ORCID 0000-0001-5318-9896

Promotor pomocniczy: dr hab. Grzegorz Koczyk

ORCID 0000-0000-5414-4689

Acknowledgements

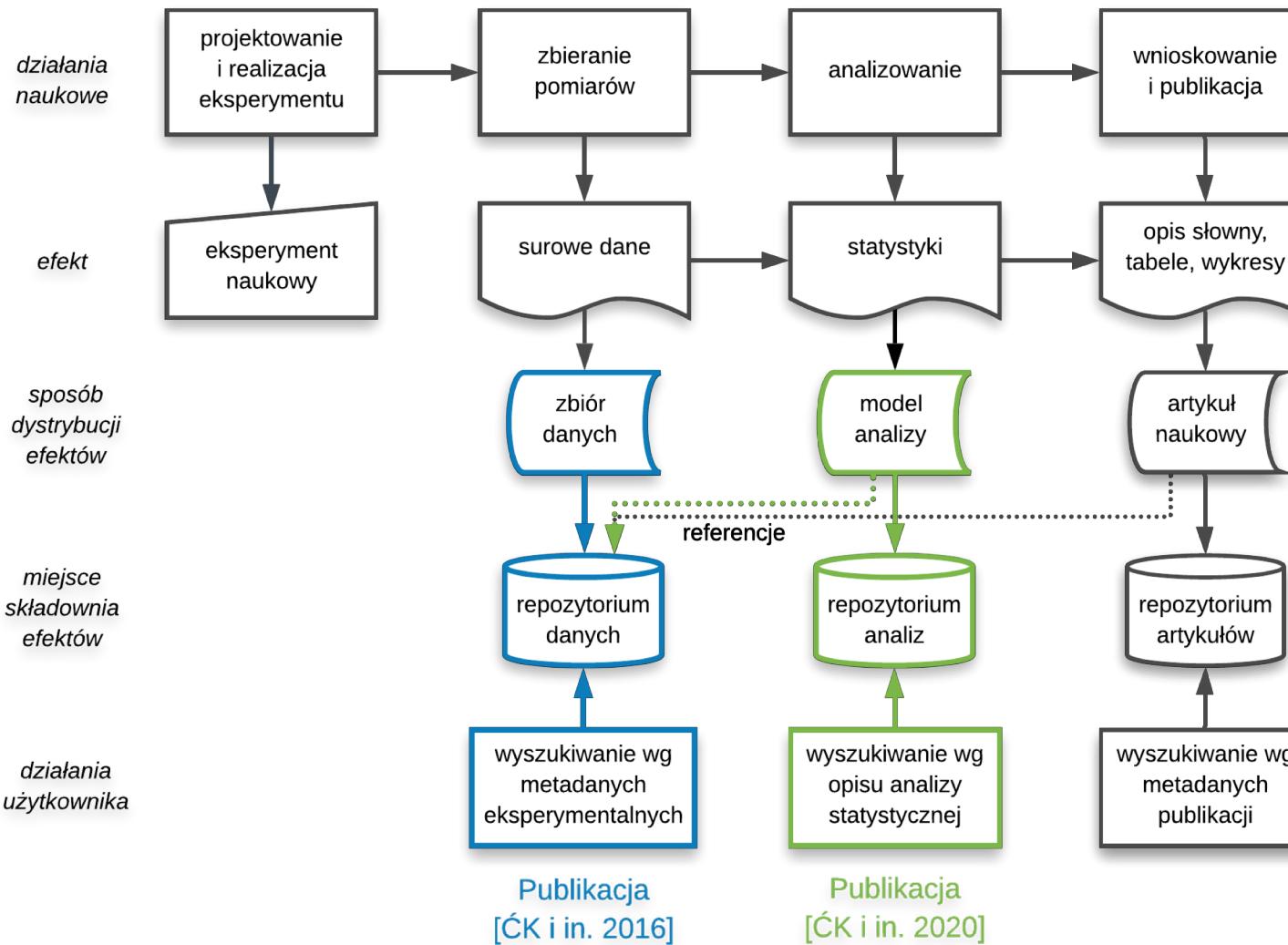
Badania zrealizowano we współpracy w ramach projektów:

- EU FP7 no. 283496
„Trans-National Infrastructure for Plant Genomic Science — TransPLANT”

- EU H2020 no. 731013
„European Plant Phenotyping Network — EPPN2020”

- NCN 2016/21/N/ST6/02358 Preludium
„Semantyczne porównywanie zbiorów danych ilościowych”

Wstęp



Cel badań:

**poprawa jakości i dostępności
roślinnych danych fenotypowych**

integralność,
weryfikacja,
odtwarzalność,
integracja,
meta-analiza,

...

Plan prezentacji

1. Wprowadzenie do tematyki i uzasadnienie badań
 - Odtwarzalność badań
 - Fenotypowanie roślin – charakterystyka danych i analiz
 - Modelowanie semantyczne – metodologia
2. Wyniki badań i publikacje
 - Dokumentowanie doświadczeń fenotypowania
 - Dokumentowanie wyników analizy statystycznej
3. Dyskusja
 - Dalsze plany rozwoju wyników
 - Szerszy kontekst i działania naukowe w społeczności fenotypowania roślin

1. Wprowadzenie

Odtwarzalność badań

Fenotypowanie roślin

Modelowanie semantyczne

Dokumentowanie i udostępnianie badań

- Otwarta Nauka [openaire.eu][otwartanauka.pl]
- Plany zarządzania danymi w EU H2020 [EC 2016] i NCN [Błocki 2019]
- Wskaźniki oceny czasopism naukowych [Nosek i in. 2015]
- Odtwarzalność badań (*reproducibility*) [Stodden 2014]
 - **Empiryczna**
 - Obliczeniowa
 - **Statystyczna**

Dane FAIR

Dane naukowe powinny być FAIR [Wilkinson i in. 2016]:

- **Findable** — odnajdywalne
- **Accessible** — dostępne na dobrze określonych warunkach
- **Interoperable** — interoperacyjne
- **Reusable** — zdatne do ponownego, poprawnego użycia



Zapewnienie własności FAIR przez spełnienia wymagań

- Technicznych — uniwersalne wymagania wobec infrastruktury do obsługi danych
- Merytorycznych — właściwie poszczególnym obszarom badań specyficzne **standardy dziedzinowe** (schematy metadanych, formaty i terminologia)

Fenotypowanie roślin jest słabo ustandaryzowane ☹

Standaryzacja danych fenotypowych

- Brak dziedzinowego standardu opisu doświadczenia fenotypowania roślin (2015)
- Brak wytycznych czasopism odnośnie publikacji zbiorów danych
- Brak centralnych, referencyjnych, dedykowanych repozytoriów danych fenotypowych
 - Por. zasoby EMBL-EBI [ebi.ac.uk/services], NCBI [ncbi.nlm.nih.gov]
- Rozproszenie uzasadnione
 - Różne pochodzenie: instytucje naukowe, badania rejestrów, hodowla roślin, rolnictwo
 - Duża różnorodność typów doświadczeń i pomiarów
 - Duże rozmiary
- Brak właściwego opisu danych utrudniają ich wykorzystanie
 - lokalne bazy danych -> różne schematy i formaty danych, nazewnictwo
 - repozytoria ogólnego przeznaczenia (np. Zenodo)
 - czaso- i kosztochłonne doświadczenia
 - cenne, unikalne obserwacje

The screenshot shows the NCBI homepage with a sidebar menu on the right. The menu includes links such as 'All Databases', 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The 'Resource List (A-Z)' link is highlighted with a blue arrow.

The screenshot shows the EMBL-EBI Services page. At the top, there are links for 'EMBL-EBI' and 'Services'. Below this, a section titled 'Browse by type' contains a 3x3 grid of categories with icons and labels: DNA & RNA (xx), Gene Expression (oculars), Proteins (grid), Structures (atom), Systems (circular arrows), Chemical biology (starburst), Ontologies (cube), Literature (book), and Cross domain (camera).

Fenotypowanie roślin

Fenotypowanie — **badanie zespołu obserwowańnych cech (fenotypu) organizmu:**

- agronomiczne (związane z uprawą i plonowaniem), morfologiczne (budowa), fenologiczne (przebieg cyklu życia), fizjologiczne (przebieg procesów życiowych), biochemicalne (ekspresja genów i białek, zawartość metabolitów), ...
- Doświadczenia roślinne
 - wybrane **genotypy roślin**
 - w określonym **środowisku**
 - z wykorzystaniem **protokołów eksperymentalnych**, dla zadanych **kombinacji czynników eksperymentalnych**
- Proces obserwacji cech fenotypowych i środowiskowych
 - za pomocą różnych narzędzi i sensorów
 - charakter ilościowy, jakościowy lub obrazowy
 - dla jednostek eksperymentalnych o różnej granularności
 - np. całe pole, poletko, grupa roślin, pojedyncza roślina, organ, tkanka, ...
 - o różnej rozdzielczości czasowej



EMPHASIS

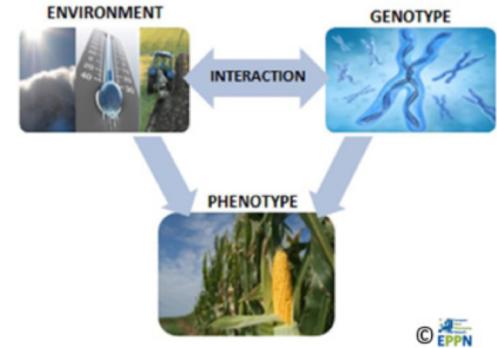


EPPN 2020



Analiza doświadczeń roślinnych

- Rodzaje analizy:
 - cech fenotypowych
 - analiza wariancji i korelacji, analiza regresji, analiza krzywych wzrostu
 - ... i cech środowiska
 - analiza interakcji genotypowo-środowiskowej $G \times E$
 - ... i informacji o genomie, np. markerów genetycznych
 - mapowanie loci cech ilościowych (QTL)
 - całogenomowa analiza asocjacyjna (GWAS)
 - Analizy integracyjne

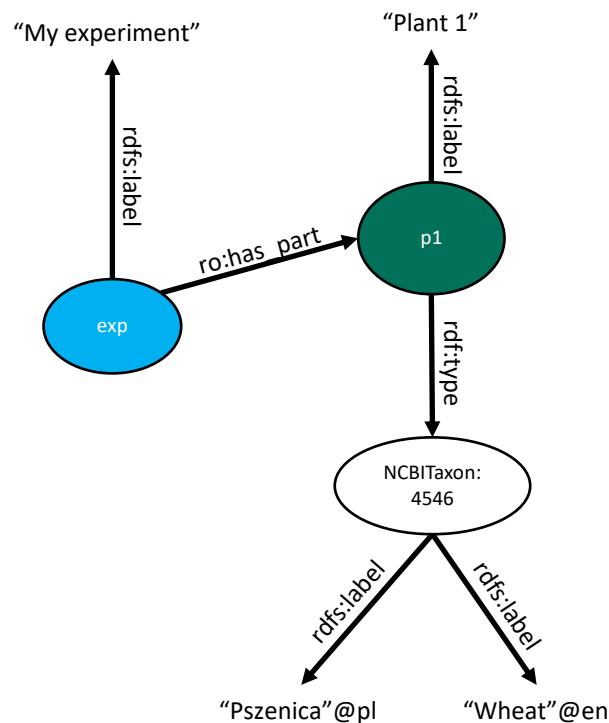


	Cechy Fenotypowe	Środowiskowe	Genotypowe i inne
Obiekty			

Liniowe modele mieszane (LMM)

- Weryfikacja hipotez badawczych
 - istotność badanych zależności
 - Estymacja parametrów i funkcji parametrycznych
 - np. wartości oczekiwane dla kombinacji czynników eksperymentalnych, kontrasty, odziedziczalność

Modelowanie semantyczne danych FAIR



Reprezentacja wiedzy poprzez **formalną definicję znaczenia (semantyki)** informacji w odniesieniu do zrozumiałego kontekstu (ontologii), realizująca ideę Sieci Semantycznej [Berners-Lee i in. 2001]

- **struktura grafowa** dla opisu różnorodnych informacji
 - Resource Description Framework (W3C RDF)
 - fakty (triples) <object> <predicate> <subject> .
 - identyfikatory IRI
- **anotacja semantyczna — klasyfikacja pojęć względem ontologii**
 - Web Ontology Language (W3C OWL)
 - sformalizowane zasoby wiedzy dziedzinowej

```
<exp> rdfs:label "My Experiment" .
<exp> ro:has_part <p1> .
<p1> rdfs:label "Plant 1" .
<p1> rdf:type <NCBITaxon:4565> .
<NCBITaxon:4565> rdfs:label "Wheat"@en .
<NCBITaxon:4565> rdfs:label "Pszenica"@pl .
```

Modelowanie semantyczne – anotacja

“Length of stem”
“Długość pędu”
“SL”
“Stem length”
“Wysokość”
“S.Len”

Plant Trait Ontology TO

stem length

http://purl.obolibrary.org/obo/TO_0000576

A stem morphology trait (TO:0000361) which is the length of the stem (PO:0009047).
<http://browser.planteome.org/amigo/term/PO:0009047>

comment
Often measured from soil surface to highest point on stem.
Refer to length (PATO:0000122): A 1-D extent quality
which is equal to the distance between two points.

has related synonym
culm height (related), STEMLG (related), CULMLG
(related), culm length (related), stem height (related), CmL
(related)

id
TO:0000576

database cross reference

- TO_GIT:157
- PATO:0000122

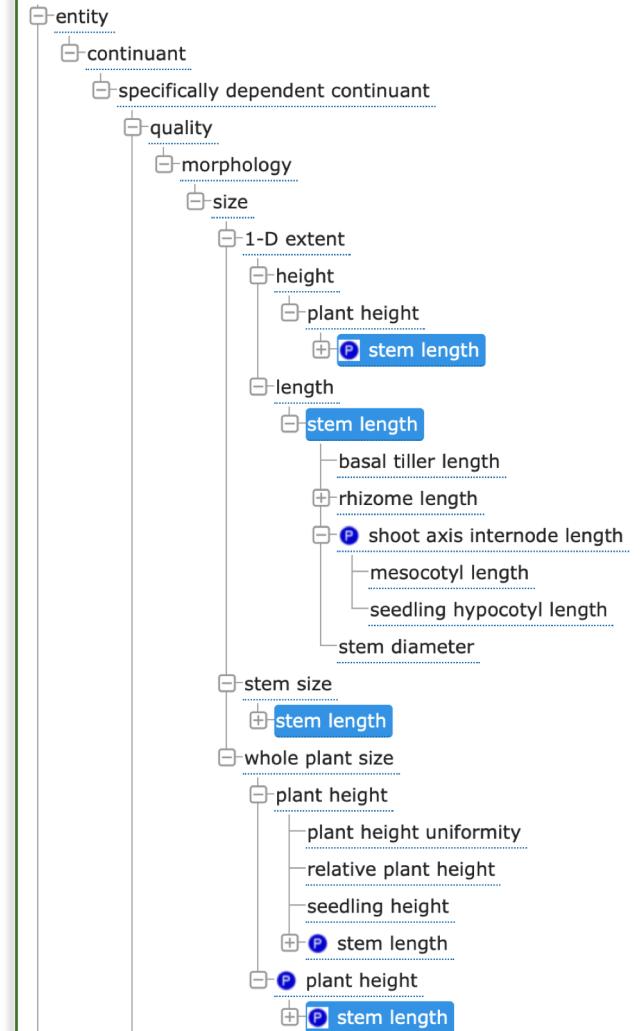
Subclass of:

- stem size
- length
- *part_of some* plant height

Related from:

part of

- shoot axis internode length

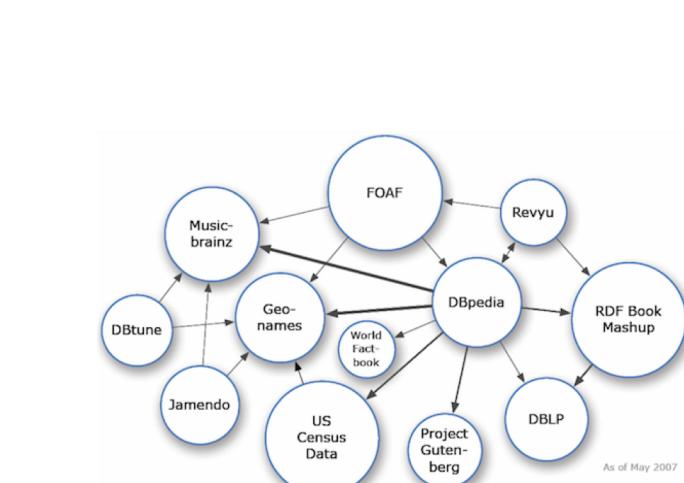


Ontology Lookup Service: ebi.ac.uk/ols

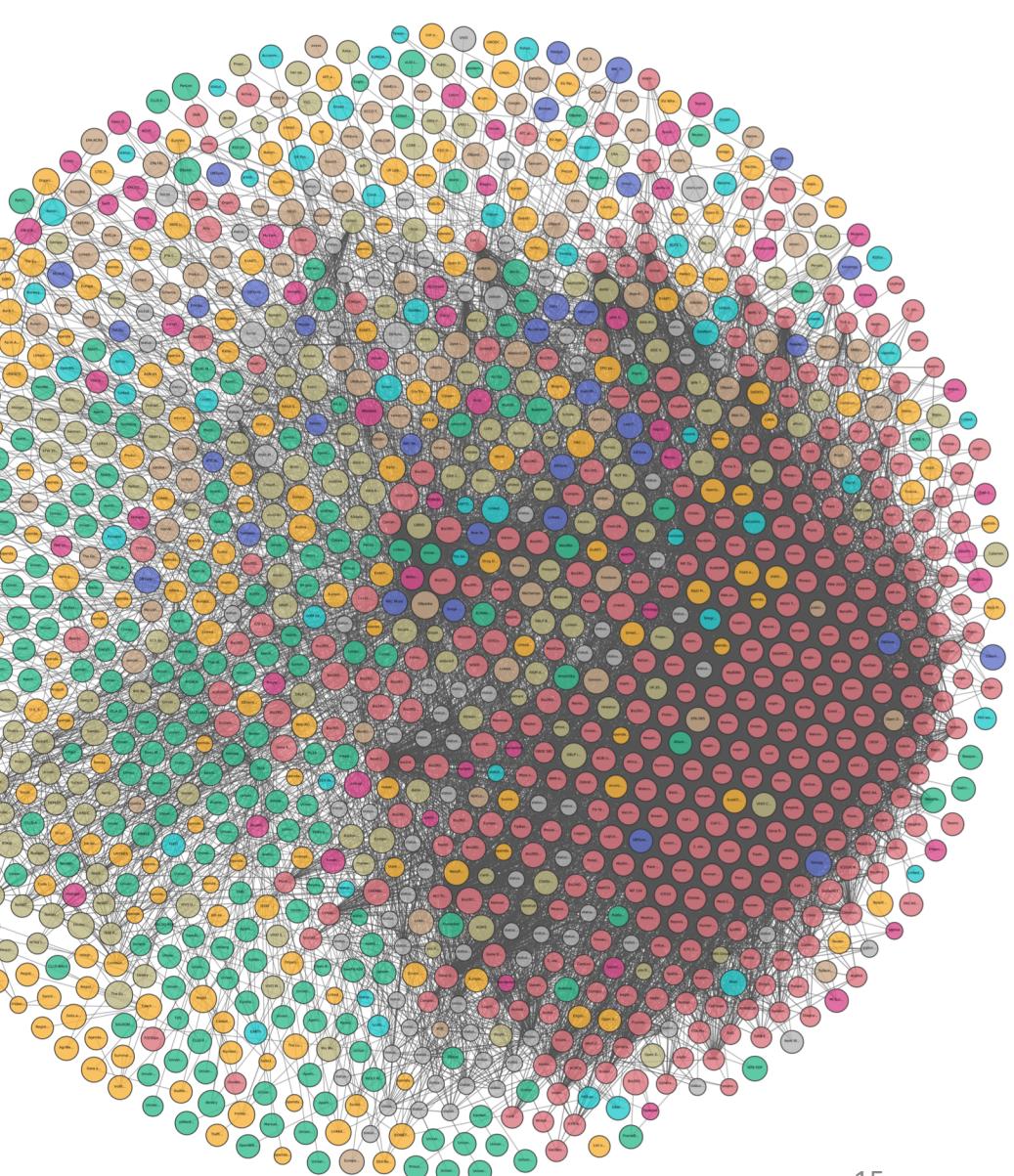
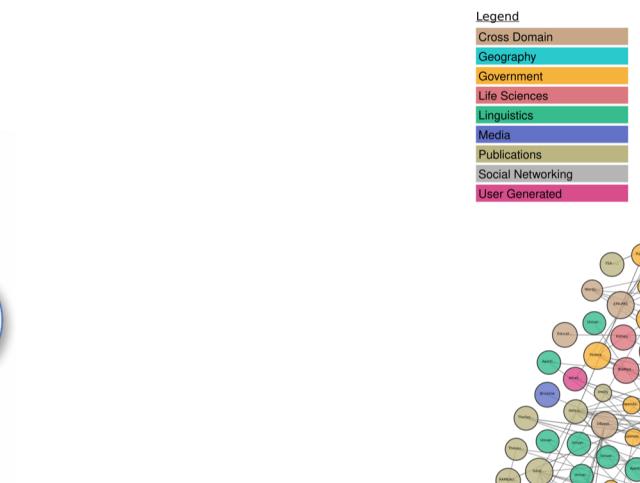
Ontologie dziedzinowe:

- Organism Taxonomy
- Gene Ontology (GO)
- Plant Ontology (PO)
- Environment Ontology (EO)
- Plant Trait Ontology (TO)
- Ontology of Biomedical Investigations (OBI)
- Experimental Factor Ontology (EFO)
- Chemical Entities of Biochemical Interest (ChEBI) ontology
- ...
- Statistics Ontology (STATO)

Linked Open Data Cloud



2007 (12)



2009 (89)

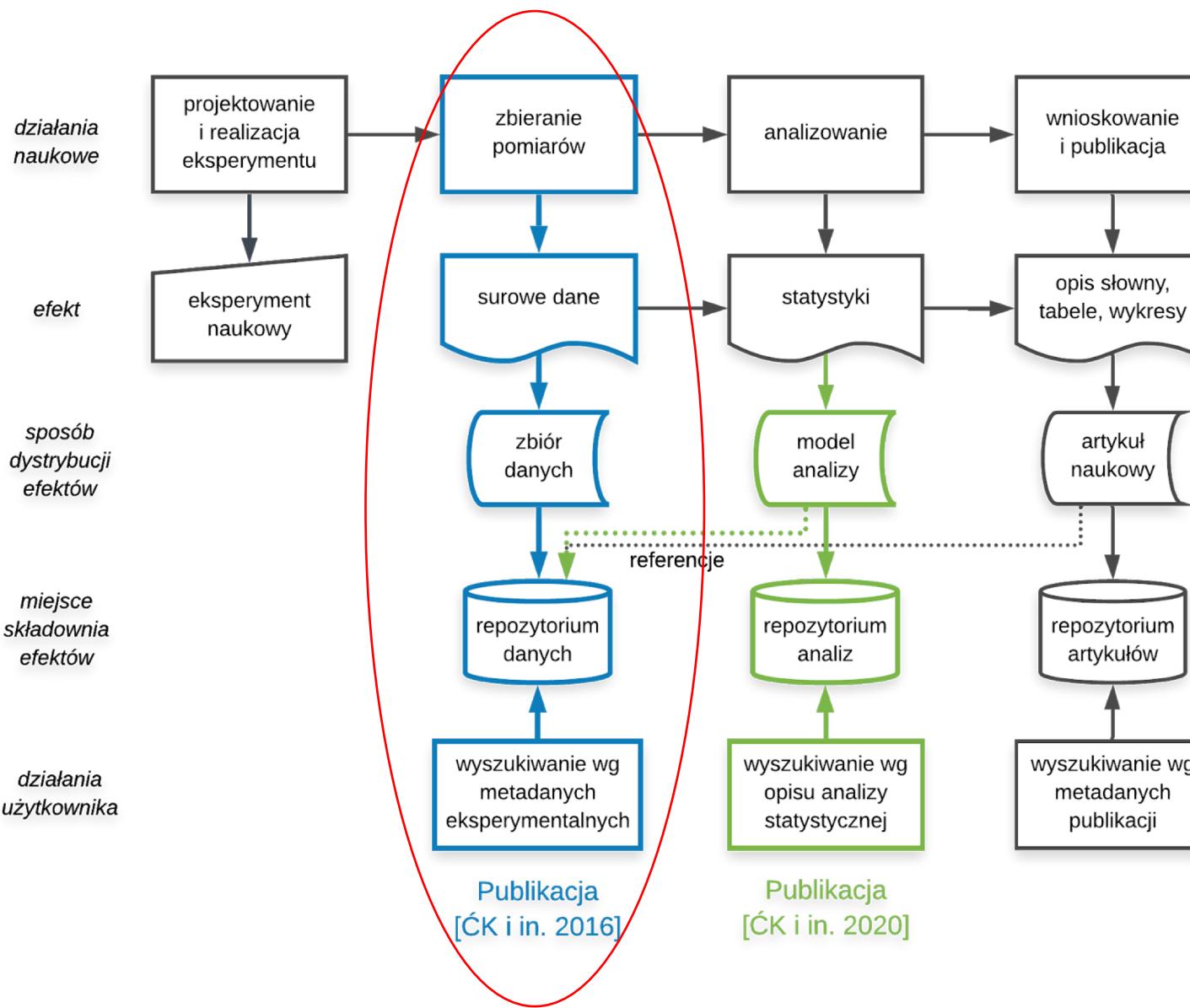
2020 (1260)

2. Wyniki badań

Modelowanie doświadczeń

Modelowanie wyników analizy statystycznej doświadczeń

Modelowanie doświadczeń



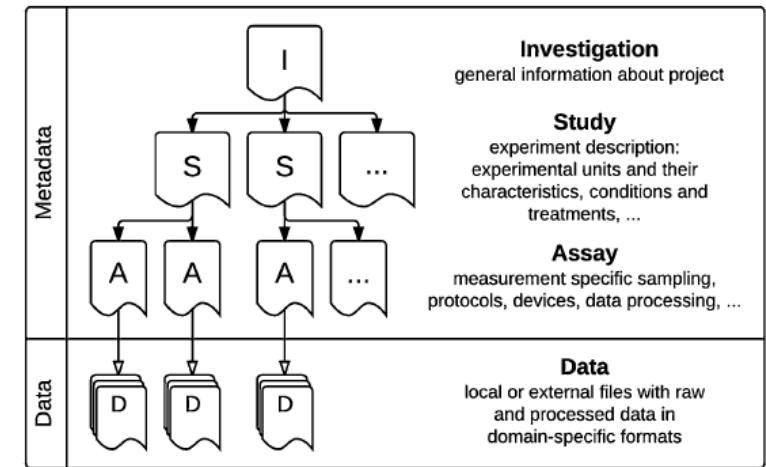
Cel badań:
poprawa jakości i dostępności
danych naukowych pochodzących
z eksperymentów **fenotypowania**
roślin poprzez zastosowanie do ich
opisu **metod semantycznych**

-> Dokumentowanie danych
i przebiegu doświadczeń

Modelowanie doświadczeń — wyniki

Standaryzacja opisu doświadczeń fenotypowania

- Zakres metadanych:
specyfikacja listy niezbędnych atrybutów doświadczeń (*checklist*)
- Terminologia:
ontologie do anotacji wartości atrybutów
- Ustrukturyzowanie:
model ISA, formaty: RDF/JSON, TAB



Model ISA [Rocca-Serra i in. 2010]

MIAPPE: Minimum Information About a Plant Phenotyping Experiment

- [FAIRsharing.org/bsg-s000543]
- miappe.org



Modelowanie doświadczeń – publikacje

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., van Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.-A., Scholz, U., van Schriek, M., Seren, Ü., Usadel, B., Weise, S., Kersey, P., & Krajewski, P. (2016). 'Measures for interoperability of phenotypic data: minimum information requirements and formatting'. *Plant Methods* (12)1, 44.

<http://dx.doi.org/10.1186/s13007-016-0144-4>

MNiSW-2016: 40 pkt.; JIF-2016: 3,510

Badania wstępne:

Krajewski, P., Chen, D., **Ćwiek, H.**, van Dijk, A. D. J., Fiorani, F., Kersey, P., Klukas, C., Lange, M., Markiewicz, A., Nap, J. P., van Oeveren, J., Pommier, C., Scholz, U., van Schriek, M., Usadel, B., & Weise, S. (2015). 'Towards recommendations for metadata and data handling in plant phenotyping', *Journal of Experimental Botany* 66(18), 5417-5427.

<https://doi.org/10.1093/jxb/erv271>

MNiSW-2015: 45 pkt.; JIF-2015: 5,677

Ćwiek-Kupczyńska, H. (2018). 'Striving for Semantics of Plant Phenotyping Data' w González-Beltrán, A., Osborne, F., Peroni, S., & Vahdati, S. (red.) *Semantics, Analytics, Visualization. SAVE-SD 2017, SAVE-SD 2018. Lecture Notes in Computer Science*. Cham: Springer International Publishing, 161-169. http://dx.doi.org/10.1007/978-3-030-01379-0_12

Ćwiek-Kupczyńska et al. *Plant Methods* (2016) 12:44
DOI 10.1186/s13007-016-0144-4

Plant Methods

METHODOLOGY

Open Access



Measures for interoperability of phenotypic data: minimum information requirements and formatting

Hanna Ćwiek-Kupczyńska¹, Thomas Altmann², Daniel Arend², Elizabeth Arnaud³, Dijun Chen⁴, Guillaume Cornut⁵, Fabio Fiorani⁶, Wojciech Frohmberg^{1,7}, Astrid Junker², Christian Klukas⁸, Matthias Lange², Cezary Mazurek⁹, Anahita Nafissi⁶, Pascal Neveu¹⁰, Jan van Oeveren¹¹, Cyril Pommier⁵, Hendrik Poorter⁶, Philippe Rocca-Serra¹², Susanna-Assunta Sansone¹², Uwe Scholz², Marco van Schriek¹¹, Ümit Seren¹³, Björn Usadel^{6,14}, Stephan Weise², Paul Kersey¹⁵ and Paweł Krajewski^{1*}

Journal of Experimental Botany, Vol. 66, No. 18 pp. 5417–5427, 2015
doi:10.1093/jxb/erv271 Advance Access publication 4 June 2015



OPINION PAPER

Towards recommendations for metadata and data handling in plant phenotyping

Paweł Krajewski^{1,*}, Dijun Chen¹, Hanna Ćwiek¹, Aalt D.J. van Dijk³, Fabio Fiorani⁶, Paul Kersey⁵, Christian Klukas⁸, Matthias Lange², Augustyn Markiewicz⁸, Jan Peter Nap³, Jan van Oeveren¹¹, Cyril Pommier⁵, Uwe Scholz², Marco van Schriek¹¹, Björn Usadel^{6,14} and Stephan Weise²

¹ Institute of Plant Genetics, Polish Academy of Sciences, ul. Strzeszyńska 34, Poznań, Poland

² Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Corrensstrasse 3, D-06466 Stadt Seeland, Germany

³ Applied Bioinformatics, Bioscience, Plant Sciences Group, Wageningen University and Research, 6706 PB Wageningen, The Netherlands

⁴ Forschungszentrum Jülich, IBG-2 Plant Sciences, Jülich, Germany

⁵ The European Molecular Biology Laboratory-The European Bioinformatics Institute, Wellcome Trust Genome Campus CB10 1SD, UK

⁶ Poznań University of Life Sciences, ul. Wojska Polskiego 28, Poznań, Poland

⁷ Keygene N.V., Agro Business Park 90, 6708 PW Wageningen, The Netherlands

⁸ INRA-URGI, Route de Saint Cyr, Versailles, France

⁹ RWTH Aachen, Werner Weg 3, Institute of Biology I, Aachen, Germany

* To whom correspondence should be addressed. E-mail: pkra@igr.poznan.pl
All authors contributed equally to this work.

Received 23 December 2014; Revised 4 May 2015; Accepted 11 May 2015

Editor: Tracy Lawson

Abstract

Recent methodological developments in plant phenotyping, as well as the growth of plant science and breeding, are resulting in a fast accumulation of multidimensional data. Addressing the challenges of managing and publishing these data, we inspect the case of plant phenotyping data publishing. We discuss how the publishers could foster advancements in the field of plant research and data analysis methods by warranting good quality phenotypic data with foreseeable semantics.

when accurately analysed and linked life. As phenotyping is a field of research can be fostered by reusing and combinability and interoperability, is possible of metadata. So far there have been no data exchange and reuse.

information about plant phenotyping and its formatting. We provide a document which specifies what information about

Striving for Semantics of Plant Phenotyping Data

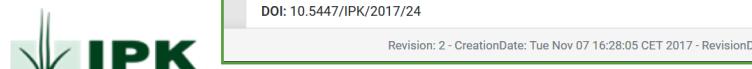
Hanna Ćwiek-Kupczyńska^(✉)

Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland
hckw@igr.poznan.pl



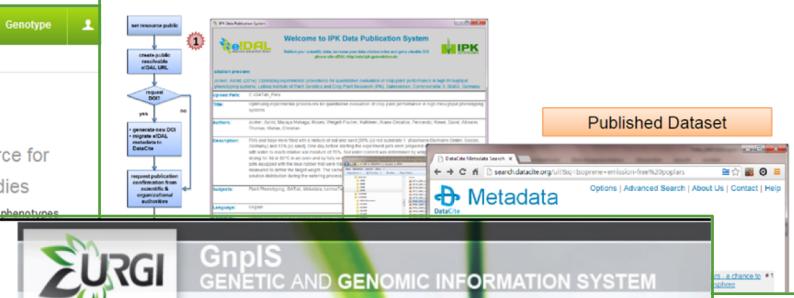
Modelowanie doświadczeń — wdrożenia

The screenshot shows the 'plantpheno' database interface. At the top, it displays 'Public datasets displayed: 9'. Below this are three dataset cards: 'Polapgen drought investigation - field PDO', 'Mapping of Quantitative Trait Loci for Traits linked to Fusarium Head Blight Symptoms Evaluation In Barley RILs', and 'POLAPGEN-BD integrative analysis demo'. On the left, there is a sidebar with 'Requirements | Funding' and a 'Search' button. Below the search is a 'Filters' section with dropdown menus for 'Trait', 'Characteristics', 'Assay Type', 'Organism', 'Infraspecific name', 'Seed origin', and 'Study type'. The main content area shows a card for 'Atwell et al., Nature 2010'.



Data Publication for DPPN Phenotype Data

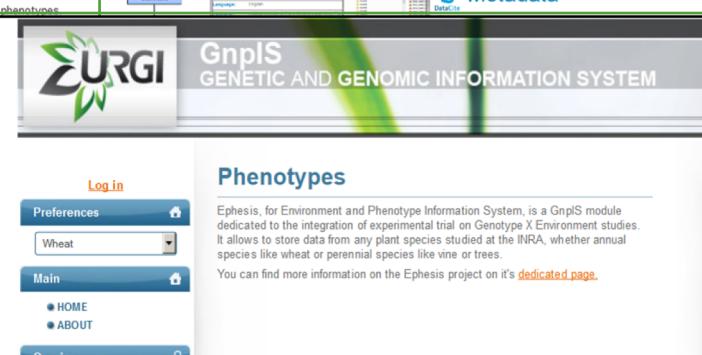
e!DAL data publication infrastructure



Welcome to
GWA-Portal Resource for
phenotypes and GWAS studies

Users can interactively browse and view public phenotypes.
Logged-in users can create studies, upload phenotypic data,
analysis using different methods on different genotypes
share the data with other users.

Take a tour



The screenshot shows the (GIGA)n SCIENCE journal website. At the top right are 'Sign In ▾' and 'Register' buttons. The header includes the journal logo '(GIGA)n SCIENCE' and navigation links for 'Articles', 'Submit ▾', 'Alerts', and 'About ▾'. The main content area features an article titled 'Predicting plant biomass accumulation from image-derived parameters' by Dijun Chen, Rongli Shi, Jean-Michel Pape, Kerstin Neumann, Daniel Arend, Andreas Graner, Ming Chen, and Christian Klukas. The article is from Volume 7, Issue 2, February 2018. The BGI logo is visible in the top right corner.



Rekomendacje RDA Wheat
Interoperability WG
odnośnie standaryzacji
danych doświadczalnych



Rozwój interfejsu programowania aplikacji Breeding API
[BrAPI.org] w oparciu o MIAPPE



Element współpracy europejskich infrastruktur
ESFRI ELIXIR i ESFRI EMPHASIS
dla danych fenotypowych

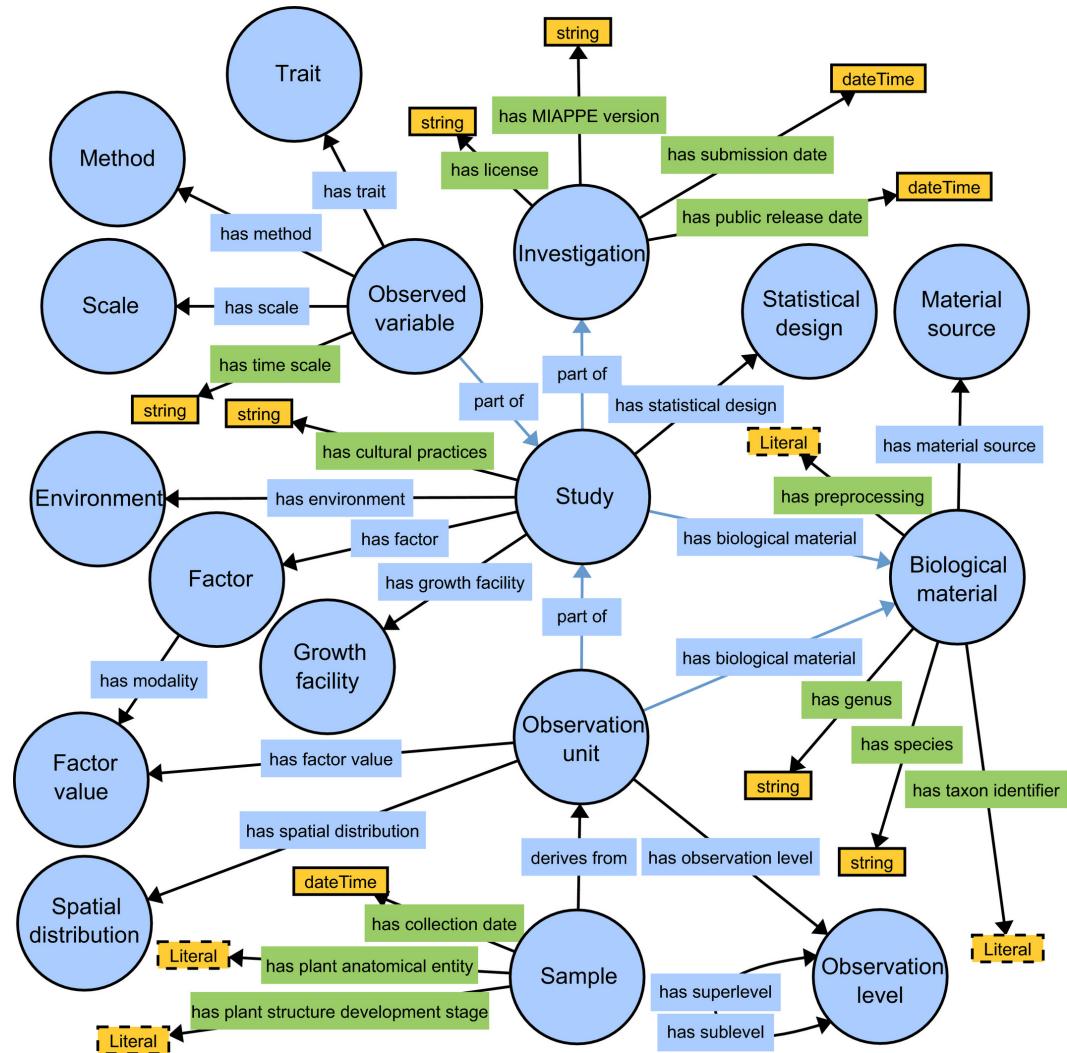
Modelowanie doświadczeń – kontynuacja badań

- Standard rozwijany społecznościowo jako miappe.org
- RFC → MIAPPE 1.1
- **Model semantyczny MIAPPE** [purl.org/ppeo.owl]
Plant Phenotyping Experiment Ontology (PPEO)
- Papoutsoglou, E. A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., Coppens, F., Cornut, G., Costa, B. V., Ćwiek-Kupczyńska, H., Drosbeke, B., Finkers, R., Gruden, K., Junker, A., King, G. J., Krajewski, P., Lange, M., Laporte, M.-A., Michotey, C., Oppermann, M., Ostler, R., Poorter, H., Ramírez-Gonzalez, R., Ramšak, Ž., Reif, J. C., Rocca-Serra, P., Sansone, S.-A., Scholz, U., Tardieu, F., Uauy, C., Usadel, B., Visser, R. G. F., Weise, S., Kersey, P. J., Miguel, C. M., Adam-Blondon, A.-F., & Pommier, C. (2020). 'Enabling reusability of plant phenomic datasets with MIAPPE 1.1', *New Phytologist* 227, 260-273. <https://doi.org/10.1111/nph.16544>
MNiSW-2019: 140 pkt.; JIF-2019: 8,512

The screenshot shows a section of a scientific article. At the top, there's a header 'Methods' with a green background. Below it, the title 'Enabling reusability of plant phenomic datasets with MIAPPE 1.1' is displayed in a large, bold, blue font. The authors' names are listed at the bottom of this section.

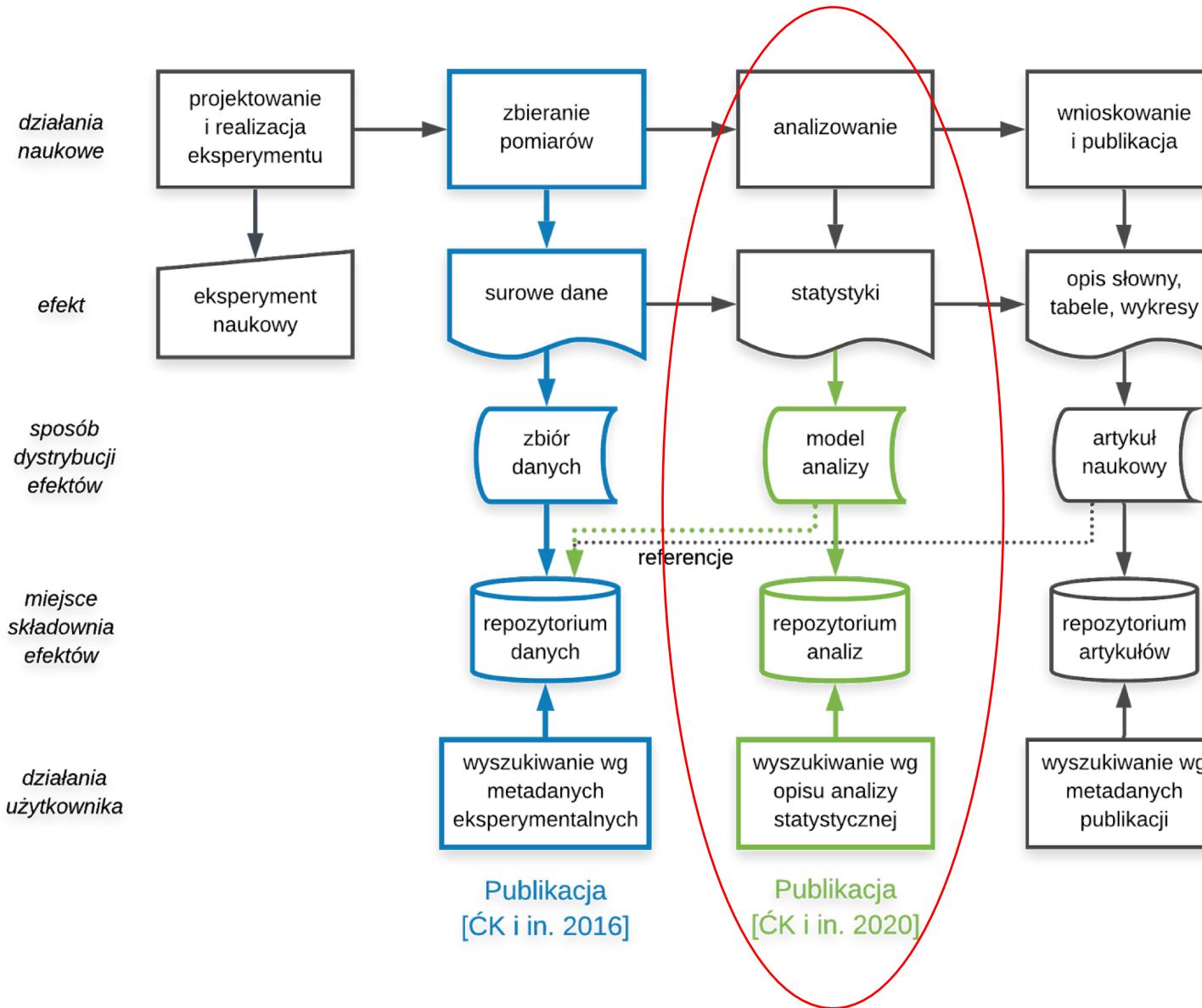
Evangelia A. Papoutsoglou¹ , Daniel Faria^{2,3} , Daniel Arend⁴ , Elizabeth Arnaud⁵ , Ioannis N. Athanasiadis⁶ , Inês Chaves^{7,8} , Frederik Coppens^{9,10} , Guillaume Cornut¹¹ , Bruno V. Costa^{7,12} , Hanna Ćwiek-Kupczyńska¹³ , Bert Drosbeke^{9,10} , Richard Finkers¹ , Kristina Gruden¹⁴ , Astrid Junker⁴ , Graham J. King¹⁵ , Paweł Krajewski¹³ , Matthias Lange⁴ , Marie-Angélique Laporte⁵ , Célia Michotey¹¹ , Markus Oppermann⁴ , Richard Ostler¹⁶ , Hendrik Poorter^{17,18} , Ricardo Ramírez-Gonzalez¹⁹ , Živa Ramšak¹⁴ , Jochen C. Reif⁴ , Philippe Rocca-Serra²⁰ , Susanna-Assunta Sansone²⁰ , Uwe Scholz⁴ , François Tardieu²¹ , Cristobal Uauy¹⁹ , Björn Usadel^{17,22} , Richard G. F. Visser¹ , Stephan Weise⁴ , Paul J. Kersey²³ , Célia M. Miguel^{7,12} , Anne-Françoise Adam-Blondon¹¹ , Cyril Pommier¹¹

¹Plant Breeding, Wageningen University & Research, PO Box 386, Wageningen 6700 AJ, the Netherlands; ²BioData.pt, Instituto Gulbenkian de Ciência, 2780-156, Oeiras, Portugal; ³INESC-ID, 1000-029, Lisboa, Portugal; ⁴Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466, Seeland, Germany; ⁵Bioversity International, Parc Scientifique Agropolis II, Montpellier Cedex 5, 34397, France; ⁶Geo-Information Science and Remote Sensing Laboratory, Wageningen University, Deveendalssteeg 3, Wageningen 6708 PB, the Netherlands; ⁷Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa (ITQB NOVA) Avenida da República, 2780-157, Oeiras, Portugal; ⁸Instituto de



Subset of the Plant Phenotyping Experiment Ontology representing the MIAPPE 1.1 data model
[Papoutsoglou i in. 2020]

Modelowanie analizy



Cel badań:

poprawa jakości i dostępności
danych naukowych pochodzących
z eksperymentów **fenotypowania**
roślin poprzez zastosowanie do ich
opisu **metod semantycznych**

-> **dokumentowanie analizy
statystycznej i wniosków
z doświadczeń**



NCN Preludium i EU H2020 "EPPN2020"

Modelowanie analizy

"Model for variable *y*,
including variable *Treatment* with *fixed effects*, ..."

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: y ~ 0 + Treatment + (1 | Block)
Data: ex1$data
```

REML criterion at convergence: 16.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.0930	-0.3092	0.0000	0.3092	1.0930

Random effects:

Groups	Name	Variance	Std.Dev.
Block	(Intercept)	0.6667	0.8165
Residual		6.1667	2.4833

Number of obs: 6, groups: Block, 2

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
TreatmentT1	5.000	1.848	2.944	2.705	0.0750 .
TreatmentT2	5.500	1.848	2.944	2.976	0.0602 .
TreatmentT3	6.000	1.848	2.944	3.246	0.0489 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	TrtmT1	TrtmT2
TreatmentT2	0.098	
TreatmentT3	0.098	0.098

Model formula with model terms with variables

Założenia:

- liniowe modele mieszane (LMM)
 - prosta struktura błędu
- $$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

Estimate of expected value of *y* for *Treatment* = *T1*

Standard error of the estimate of expected value of *y*
for *Treatment* = *T3*

Significance level of *Treatment*
for F statistics in the test of hypothesis ...

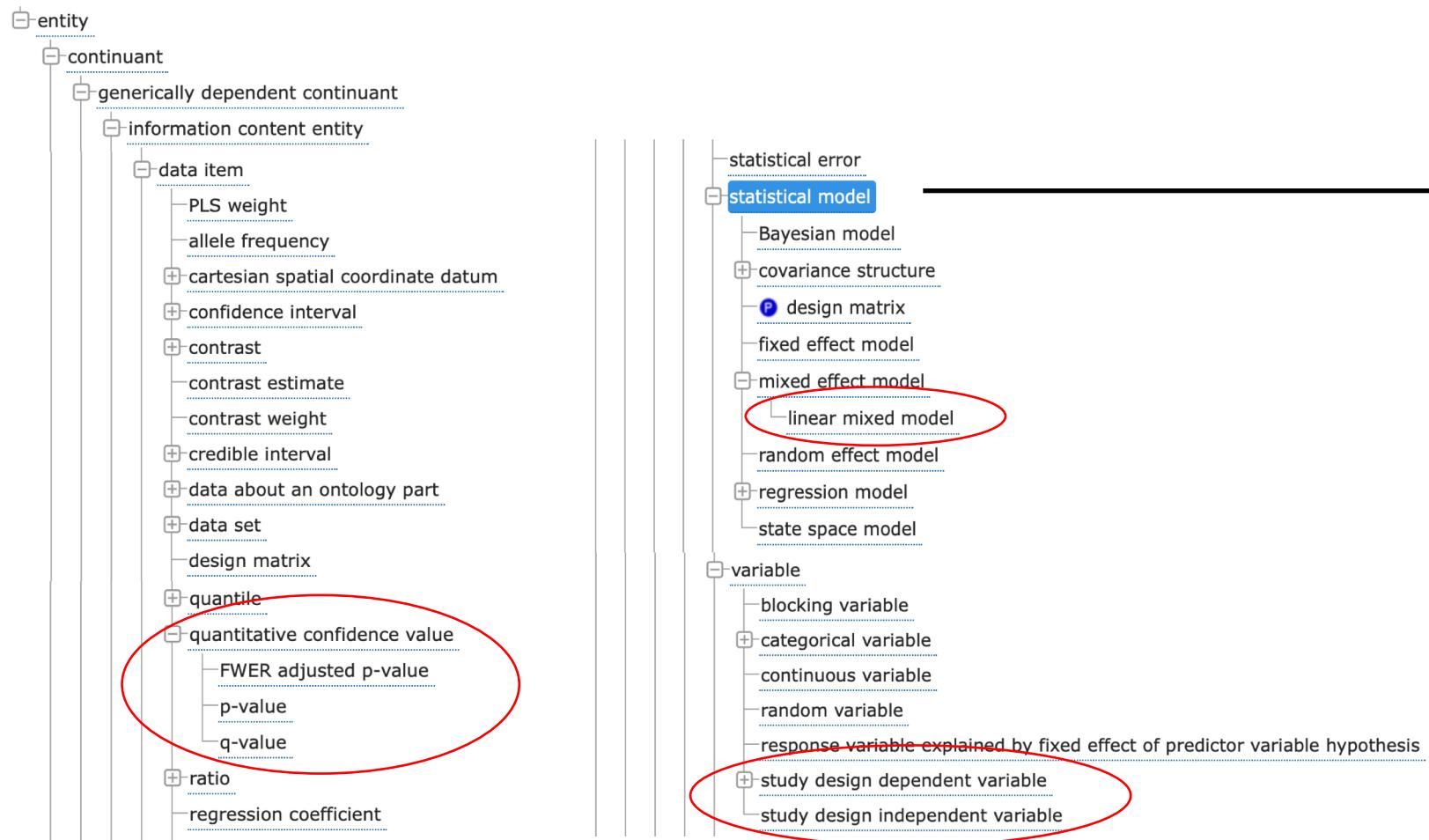
Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Treatment	138.05	46.017	3	2	7.4622	0.1205

Modelowanie analizy — anotacja

STATO: the statistical methods ontology

STATO is the statistical methods ontology. It contains concepts and properties related to statistical methods, probability distributions and other concepts related to statistical analysis, including relationships to study designs and plots.



stato-ontology.org
purl.obolibrary.org/obo/stato.owl

wyodrębniony moduł STATO-LMM:
purl.org/stato-lmm

Term relations

Subclass of:

- data item
- is denoted by some design matrix
- is_specified_input_of some model fitting
- has_part some model parameter
- is_model_for some study design dependent variable

Related from:

is declared by

- model error term
- model term

has_specified_input

- Breusch-Pagan test
- breeding value estimation

is_about

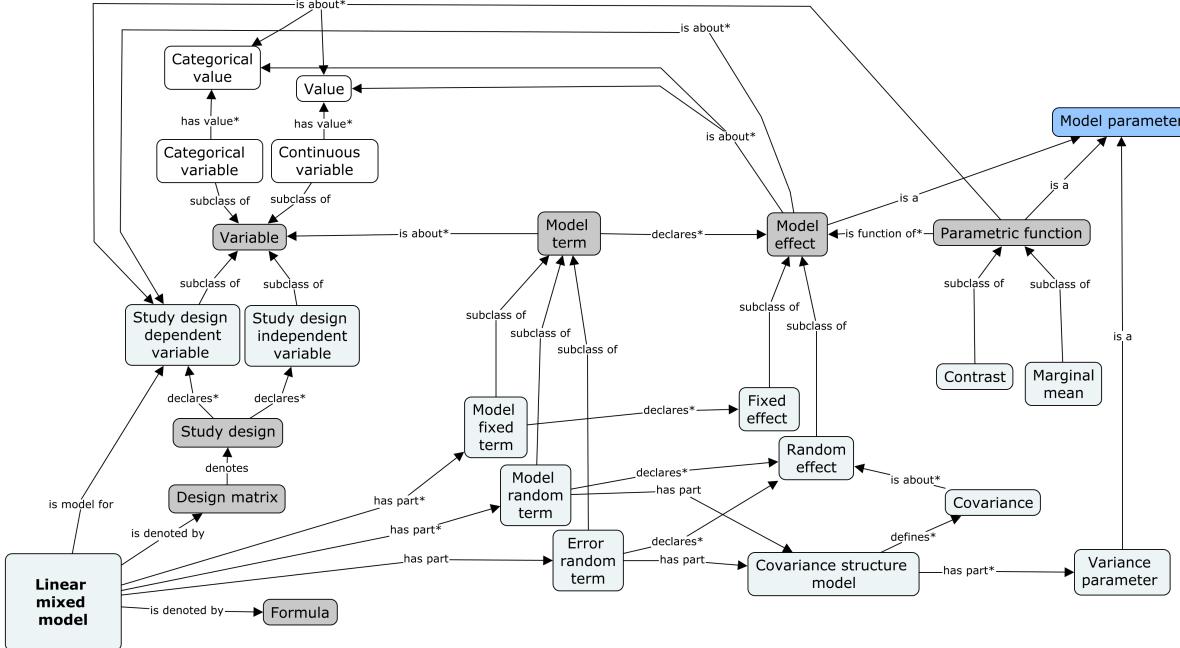
- Bayes factor
- deviance information criterion
- focused information criterion
- Akaike information criterion
- Bayesian information criterion
- deviance
- Mallows' Cp

part of

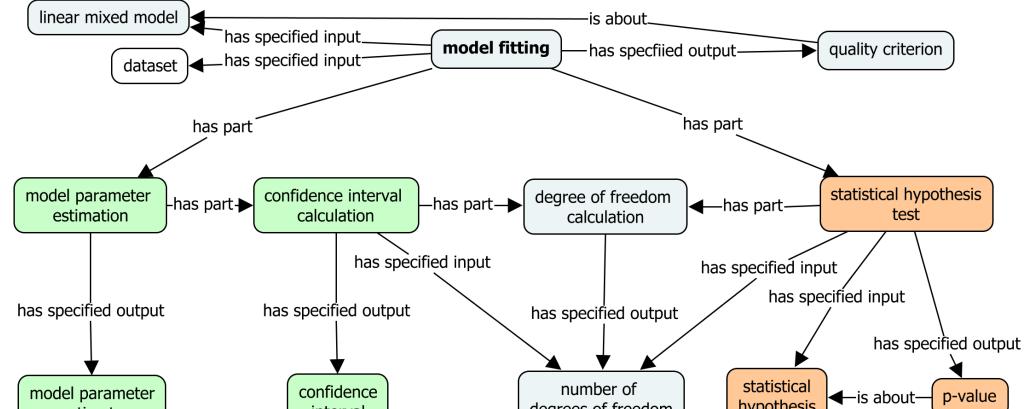
- design matrix

Modelowanie analizy – *ontology design patterns*

Deklaracja LMM:

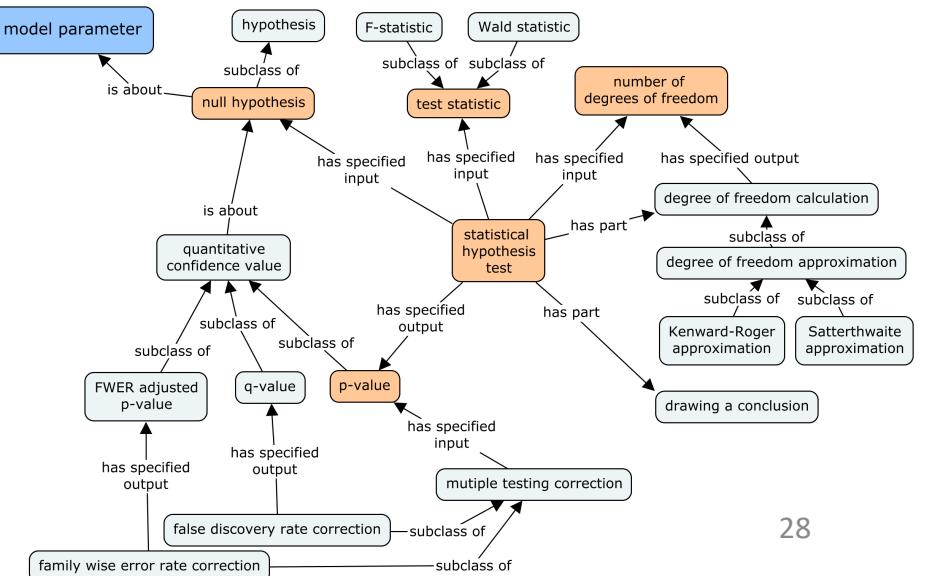
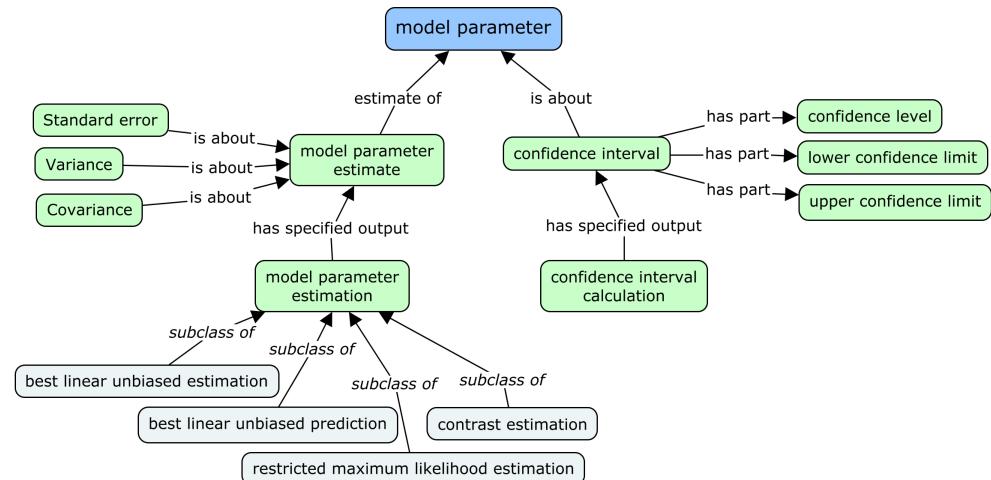


Dopasowanie LMM:



Testowanie:

Estymacja:



Modelowanie analizy — wyniki i publikacja

Semantyczny model wyników statystycznej analizy LMM

- Anotacja LMM za pomocą Statistics Ontology (STATO)
- Rozbudowa STATO o terminy związane z LMM
- Udostępnienie wyników analizy statystycznej zbiorów danych z doświadczeń fenotypowania dla zautomatyzowanego wyszukiwania zbiorów danych

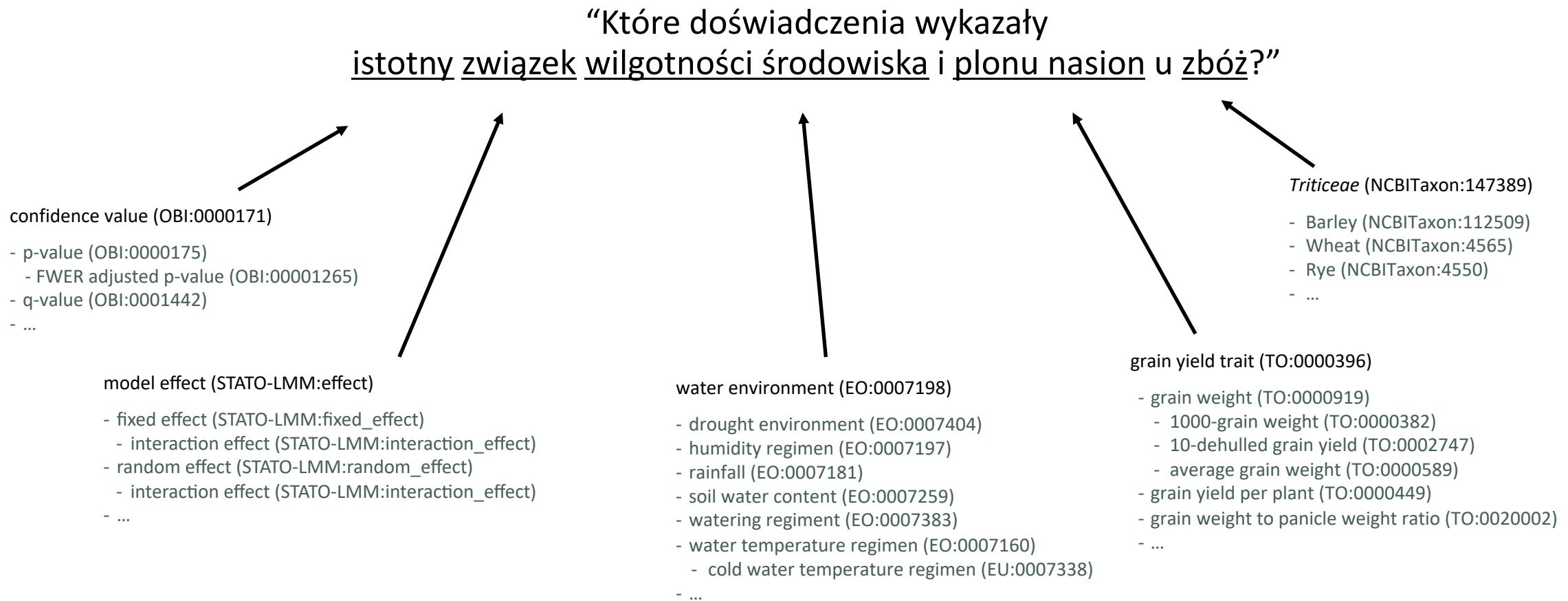
Ćwiek-Kupczyńska, H., Filipiak, K., Markiewicz, A., Rocca-Serra, P., Gonzalez-Beltran, A. N., Sansone, S.-A., Millet, E. J., van Eeuwijk, F., Ławrynowicz, A., & Krajewski, P. (2020). ‘Semantic concept schema of the linear mixed model of experimental observations’. *Scientific Data*, 7(1), 70. <https://doi.org/10.1038/s41597-020-0409-7>
MNiSW-2019: 140 pkt.; JIF-2019: 5,541

The image shows a screenshot of a journal article from the 'SCIENTIFIC DATA' journal. The article is titled 'Semantic concept schema of the linear mixed model of experimental observations'. It is an open-access article by Hanna Ćwiek-Kupczyńska, Katarzyna Filipiak, Augustyn Markiewicz, Philippe Rocca-Serra, Alejandra N. Gonzalez-Beltran, Susanna-Assunta Sansone, Emilie J. Millet, Fred van Eeuwijk, Agnieszka Ławrynowicz, and Paweł Krajewski. The abstract discusses the development of a semantic model for statistical analysis using ontologies like STATO to produce FAIR data summaries, improving the understanding and automation of statistical modeling.

Modelowanie analizy — przykłady

- Czy są dostępne doświadczenia, w których badano interakcję czynników A i B ?
- Czy wykazano gdzieś istotny związek czynnika C i cechy Y ?
- Które obserwowane cechy są wykazują istotną zmienność w doświadczeniach, w których występuje czynnik D ?
- Jakie wartości czynników E i F łączą się z maksymalnymi wartościami cechy Y ?

Modelowanie semantyczne – przykład



Modelowanie semantyczne – przykład

SPARQL Query & Update

```

significant fixed effects × + ⓘ
1 PREFIX obo: <http://purl.obolibrary.org/obo/>
2 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 PREFIX stato: <http://purl.obolibrary.org/obo/stato.owl#>

# Find significant fixed effects.
# Returns fixed effects (relative effects, contrasts, i.e. difference against
reference) of variable levels for a specified trait (here "GW_m2"), ordered by the
absolute value of its estimate, if the corresponding p-value is significant.
# Input: Trait, p-value (see below)

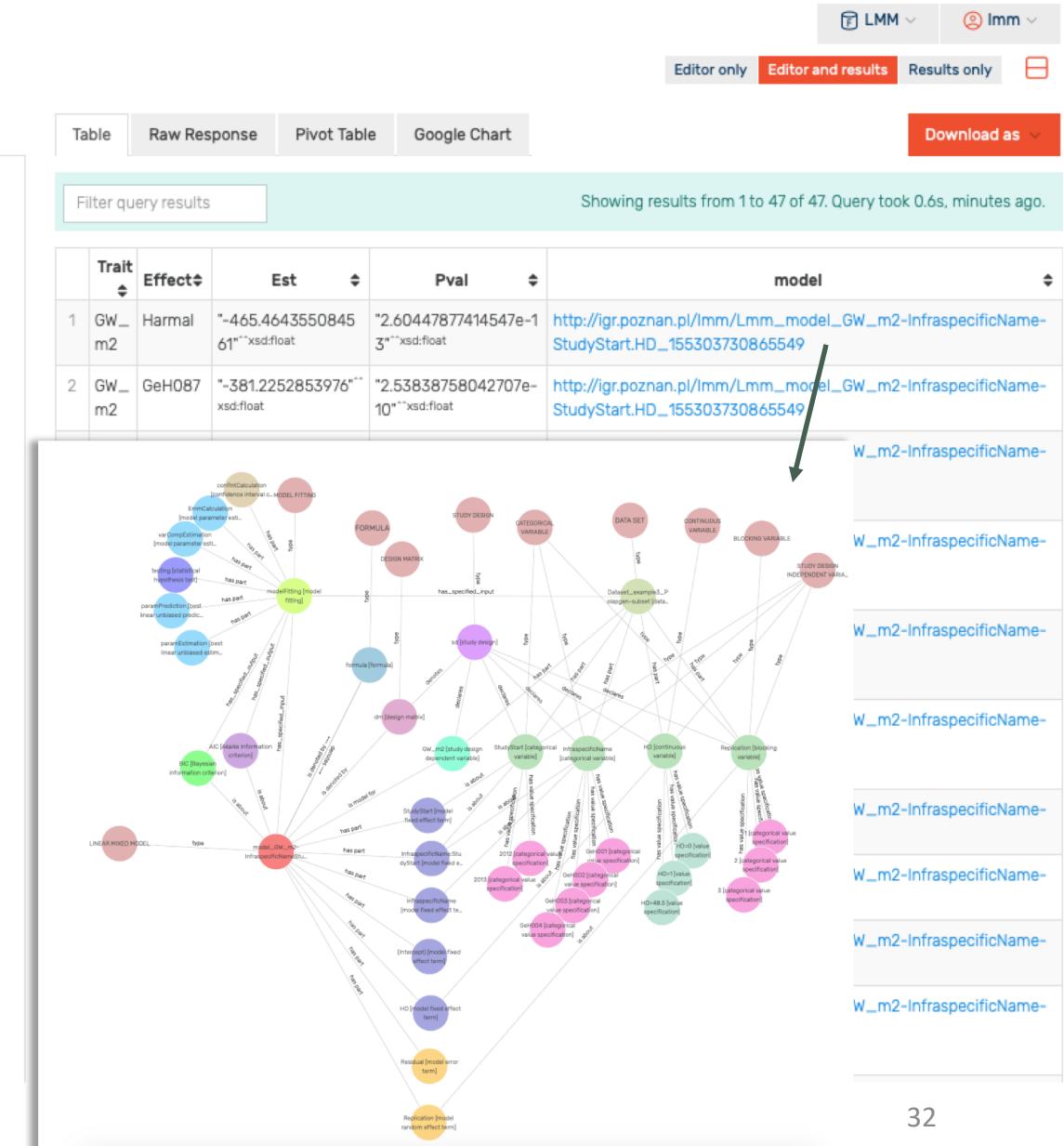
SELECT DISTINCT
    ?Trait ?Effect ?Est ?Pval ?model
WHERE {
    FILTER(regex(?Trait, "GW_m2")). ##### HERE SPECIFY THE DEPENDENT VARIABLE

    ?model rdf:type obo:STATO_0000464. #mixed effect model
    ?model obo:BFO_0000051 ?term.
    ?model stato:is_model_for/rdfs:label ?Trait.
    ?term rdfs:label ?Term.
    ?term obo:STATO_0000001 ?effect, ?effect2.
    FILTER(?effect != ?effect2). # a term has >1 effect (intention: to exclude intercept)
    ?effect rdf:type obo:STATO_0000307. #treatment contrast
    ?effect rdfs:label ?Effect.
    ?est obo:STATO_0000403 ?effect. #estimator of
    ?est rdf:value ?Est.

    ?hypo rdf:type obo:STATO_0000065. #hypothesis
    ?hypo obo:IAO_0000136 ?effect. #is about
    ?pval obo:IAO_0000136 ?hypo. #is about
    ?pval rdf:value ?Pval.

    FILTER (?Pval < 0.01). ##### HERE SPECIFY THE SIGNIFICANCE LEVEL
}
ORDER BY desc(abs(?Est))

```



Modelowanie semantyczne – metody i narzędzia

- Rozszerzenie ontologii STATO: *Protégé OWL editor*
- Ekstrakcja modułu ontologii STATO-LMM: *OntoFox & MIREOT*
- Analiza statystyczna danych za pomocą LMM: *R lme4* i *R nlme* oraz *Genstat VSNI*
- Generowanie modelu semantycznego LMM: *R* (skrypt własny) -> *.trig
- Baza danych: *GraphDB triple store & SPARQL query endpoint*

3. Dalsze badania

Dalszy rozwój i wdrażanie rozwiązań

Szerszy kontekst i działania społeczności fenotypowania roślin

Dalsze badania:

- Model doświadczenia
 - dalszy rozwój i rozszerzenie standardu MIAPPE
 - dokumentowanie specyficznych (złożonych typów) danych środowiskowych
 - dokumentowanie technologii fenotypowania (Semantic Sensor Network Ontology, SSNO)
 - integracja danych multi-omicznych
 - rozbudowa narzędzi do generowania modelu semantycznego (PPEO) doświadczeń
 - na podstawie modelu ISA i Breeding API
- Model analizy doświadczeń
 - rozszerzenie modelu
 - dla zaawansowanych struktur kowariancji, funkcji parametrycznych i modeli wielowymiarowych
 - uzupełnienie modelu o aspekt obliczeniowy (The Software Ontology, SWO)
 - rozbudowa narzędzi (skrypt R) do generowania modelu semantycznego analizy
 - na podstawie wyników LMM w pakietach R i Genstat
 - anotacja modelu za pomocą Ontology of Biological and Clinical Statistics (OBCS)
 - modelowanie semantyczne innych metod analizy danych niż LMM

Dalsze perspektywy:

- Wdrożenie
 - Stworzenie systemu gromadzenia semantycznych modeli zbiorów danych (dokumentacja doświadczeń i analiz)
 - z przyjaznym interfejsem użytkownika
 - predefiniowane zapytania w języku naturalnym > SPARQL
 - wizualizacja wyników
 - Rozszerzenie specyfikacji BrAPI o wyniki analizy statystycznej
 - Rozszerzenie i uporządkowanie ontologii STATO
 - Poprawa jakości ontologii dziedzinowych i mapowań pomiędzy nimi
 - łatwiejsza integracja różnych dziedzin

Acknowledgements



Paweł Krajewski
Monika Mokrzycka



Augustyn Markiewicz



Katarzyna Filipiak
Agnieszka Ławrynowicz
Wojciech Frohberg



Jan van Oeveren
Marco van Shriek



Paul Kersey



Björn Usadel
Fabio Fiorani
Hendrik Poorter



Uwe Scholz
Matthias Lange
Astrid Junker
Stephan Weise
Daniel Arendt



Hadi Quesneville
Cyril Pommier
Francois Tardieu
Pascal Neveu
Anne-Françoise Adam-Blondon



Célia Miguel
Inês Chaves



isatools
Susanna-Assunta Sansone
Philippe Rocca-Serra
Alejandra González-Beltrán



Magnus Nordborg
Ümit Seren



WAGENINGEN
UNIVERSITY & RESEARCH

Fred van Eeuwijk
Aalt-Jan van Dijk
Jan Peter Nap
Richard Finkers
Evangelia Papoutsoglou



Elizabeth Arnaud



Daniel Faria



Frederik Coppens



Referencje

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). 'The Semantic Web', *Scientific American* 284(5), 34-43. <https://doi.org/10.1038/scientificamerican0501-34>
- Błocki, Z. (2019). *Plany NCN w zakresie zarządzania danymi naukowymi*. [online]. https://ncn.gov.pl/sites/default/files/pliki/2019_04_03_pismo_dyrektora_NCN_zarzadzanie_danymi_naukowymi.pdf
- EC (European Commission). (2016). *Guidelines on FAIR Data Management in Horizon 2020*. [online]. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- Nosek, B. A. et al. (2015). 'Promoting an open research culture', *Science* 348(6242), 1422-1425. <https://doi.org/10.1126/science.aab2374>
- Stodden, V. (2014). 'What scientific idea is ready for retirement?', Edge.org. [Online]. <https://www.edge.org/response-detail/25340>.
- Wilkinson, M. D. et al. (2016). 'The FAIR Guiding Principles for scientific data management and stewardship.', *Scientific data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Dziękuję za uwagę