Casus data voor de wetenschap (Data science challenges)

Project titel

Meerwaarde identificeren voor data-driven consumer science door het FAIR maken WUR consumenten databases in het Europese Consumenten Data Platform.

Project leader

Karin Zimmermann

LNV contact person (PM)

Casper Zulim de Swarte

Bijdrage van het project aan de ambitie van het DDSDS-programma (DDSDS = data driven science in a digital community)

De ambitie van het DDSDS-programma is het opbouwen van expertise op het gebied van data science om met behulp van geavanceerde analytische methoden meer en nieuwe inzichten te halen uit onderzoeksgegevens. Dit project is in lijn met die ambitie en zet data-driven concepten en tools in om binnen het domein van de WUR de relaties tussen gedrag, voedsel, voeding en gezondheid van consumenten verder te onderzoeken. Daarmee wordt binnen de WUR een brug gebouwd tussen de ontwikkeling van kennis van moderne data science aan de ene kant en domeinkennis op het gebied van consumentengedrag aan de andere kant.

Introductie voor het KennisOnline informatie-portal

In dit project worden de eerste bouwstenen uitgewerkt en getoetst (proof of principle) voor een toekomst bestendige data /gegevens infrastructuur als basis voor het uitrollen van een WUR Consumer Data Platform. De technische, organisatorische en business modellen resultaten uit het Europese project RICHFIELDS zullen als uitgangspunt gelden. Expertise op gebied van consumenten onderzoek, ICT en data science en business modellen en valorisatie wordt gezamenlijk opgepakt om goede integratie en een duurzame gegevens- en kennisinfrastructuur te waarborgen. De doelstelling van dit project is om de mogelijkheid en de potentie van het FAIR maken van WUR data op het gebied van consumentengedrag te onderzoeken en toekomsite standaarden in lijn te brengen met WDCC en FNH-RI omgevingen. De ambitie van het DDSDS-programma is het opbouwen van expertise op het gebied van data science om met behulp van geavanceerde analytische methoden meer en nieuwe inzichten te halen uit onderzoeksgegevens.

Probleemdefinitie

Om binnen de WUR data science te kunnen beoefenen op het gebied van consumentengedrag is het essentieel dat onderzoeksdata snel beschikbaar zijn en in een overdraagbare (“machine readable”) vorm zodat onderzoekers gemakkelijk data kunnen delen en gebruiken. Huidig data management beperkt zich doorgaans tot het veilig opslaan en beheren van onderzoeksdata. Hiermee is de data nog niet toegankelijk voor interne of externe onderzoekers. WR heeft een eerste technisch ontwerp van een Consumer Data Platform ontwikkeld. Een stap in de goede richting, maar er is meer nodig om data science succesvol binnen consumentengedragsonderzoek toe te passen. Binnen het EU project “RICHFIELDS”, een onderdeel van de Europese Food Nutrition and Health Research Infrastructure (FNH-RI), heeft WR meegewerkt aan een conceptueel ontwerp van een Europees Consumer Data Platform (CDP) waarin de voorwaarden en eisen worden gesteld voor de realisatie waarbinnen onderzoeksdata vindbaar, toegankelijk, overdraagbaar en herbruikbaar zijn. De uitdaging is om het Consumer Data Platform van WR in lijn te brengen met de voorwaarden en eisen die het Europese Consumer Data platform uit RICHFIELDS stelt. Daarbij moet het CDP aansluiten bij de principes van FAIR data (Findable, Accessible, Interoperable, Reusable) om internationale acceptatie te realiseren. De mate van aansluiting bij FAIR principes kan via recent ontwikkelde methoden objectief bepaald worden. De realisatie ervan vindt plaats met behulp van technologieën zoals het semantisch web / linked data. Afgezien van de technische aspecten zijn hierbij de organisatorische (governance) aspecten van groot belang die partijen in staat stellen om in overeenstemming met hun (commerciële of publieke) doelstellingen data te delen.

Doelgroep en kennis

WUR onderzoekers in het domein van consumenten, voedsel, voeding en gezondheid. Deze onderzoekers hebben veel data beschikbaar, maar nog niet de middelen deze effectief te delen.

Project doelstellingen

De doelstelling van dit project is om de mogelijkheid en de potentie van het FAIR maken van WUR data op het gebied van consumentengedrag te onderzoeken en toekomsite standaarden in lijn te brengen met WDCC en FNH-RI omgevingen. Specifiek zal de focus liggen op:

- Training van consumentenwetenschappers op het gebied van data management en toepassen FAIR data.
- Standaardisatie en koppelen van specifieke datasets door middel van FAIR toepassingen.
- Maken van het CDP voor het delen, inzichtelijk maken en visualiseren van data.
- Inschatten van de mogelijkheden tot valorisatie van open en gekoppelde data.

Wetenschappelijke relevantie

Meerwaarde creëren door bestaande data te gebruiken of huidige datasets aan te vullen gebeurt weinig in het domein van consumentenonderzoek. Om deze meerwaarde te kunnen creëren moet aan een aantal randvoorwaarden voldaan worden: het beperken van het zoeken naar (consumenten) data en het verminderen van eigen interpretaties over relaties tussen de verschillende data sets. De uitdaging is dat de datasets heel verschillend zijn omdat er veel verschillende methodes gebruikt worden (andere vragenlijsten, andere aanpakken enzovoorts). Datasets kunnen op meerdere manieren aan elkaar gerelateerd worden. Momenteel gebeurt dit veelal door handwerk en beschrijving van de dataset in de vorm van metadata. Deze eerste stap, het automatiseren van het verbinden van onderzoeksdata maakt de weg vrij voor de volgende stappen, zoals ontwikkelen van specifieke standaarden, tools en algoritmes.

Activiteiten

In het overzicht hieronder word een beeld van de activiteiten weergegeven. De integratie van de activiteiten is belangrijk om vanaf het begin van het project de samenwerking, kennisuitwisseling en eenduidigheid in op te leveren producten te waarborgen. Tevens borgt deze aanpak de goede documentatie van het proces, tussen resultaten en vervolgstappen zodat de behoeftes van de gebruikers, technische eisen, data eisen en stappen gericht op ontsluiting en kennisvalorisatie met elkaar in balans blijven. De focus ligt dus op outputs die gezamenlijk worden opgesteld.

	Management	Data Science en valorisatie	Databases	Consumer Datat Platform	Gezamenlijke output
Voorbereidende fase	• Uitschrijven van project voorstel; projectteam • Interactie met WDCC, ELIXIR en FNH-RI				• Uitgewerkt projectvoorstel • Projectteam Kick off
Analysefase	• Workshop FAIR data organiseren • Projectmanagement	• Vaststellen uitgangsvraag • Requirements eenduidig in kaart brengen • Waarborgen link met business value proposition • Protocol databebruik	• Verzamelen case specifieke datasets	• Schets business-, information-systems-, en tech infr-architectuur van prototype op basis van Richfields conceptueel ontwerp • Schets huidge CDS	• Architectuurvisie WUR-specifiek CDP (volgens FAIR principes) • Visiedocument data toegang en data delen
Ontwerpfase	• Projectmanagement	• Organiseren formele gebruikersreview en documenteren feedback • Informed consent mogelijkheden	• Vertegenwoordigen onderzoekers in ontwikkelproces en formele gebruikersreview van ontwerp	• Gap analyse WR CDS en gewenste CDS • Selecteren refmodellen en standaarden • Definiëren componenten voor ontwikkeling	• Gedetailleerd ontwerp (samenhang business, information systems en technologische infrastructuur)
Ontwikkelfase	• Projectmanagement	• Review fit ontwikkelde componenten met business value proposition (perspectief kennisproduct) • Discussie invulling valorisatiestrategie voor deze casus	• Gebruiksinstructies testgroep • Passend maken datasets (FAIR principes)	• Ontwikkelen componenten inclusief visausatiemogelijkheden • Testen (litechnisch) van componenten en performance	• Werkend prototype inclusief gebruikershandleiding • Valorisatiestrategie
Test& Evaluatiefase	• Projectmanagement • Interactie met WDCC, ELIXIR en FNH-RI • Ontwikkelen visie vervolgstappen	• Ontwikkelen en toetsen valorisatiestrategie	• Testen van de ontwikkelde componenten met de passend gemaakte datasets • Beschrijving van innovaitepotentieel tooling	• Oplossen bugs • Ondersteunen gebruikerstest	• Test- en evaluatierapport • Visiedocument vervolgstappen

Op basis van EU project RICHFIELDS ligt er een technisch, organisatorisch en financieringsontwerp voor het Europese CDP. Het is van strategisch en praktisch belang om het WUR CDP hierop te laten aansluiten.

Er zijn een aantal data sets beschikbaar, die eigendom zijn van WEcR:

- FoodProfiler biedt de mogelijkheid om in een internationale context data te verzamelen om onderzoek te doen naar consumentengedrag in relatie tot voedselconsumptiepatronen. De neartime registratie van de afgelopen 2 uur, de langdurige metingen en grote groepen consumenten leveren rijke en betrouwbare inzichten. De data van de FoodProfiler volgt een EMA structuur, waarbij stukjes data van verschillende personen geaggregeerd wordt.
- ISAFRUIT data set: representatieve data sets (n=2,803), SPSS, 2008 (fruit, consumenten gedrag (extended), zelf gerapporteerde voedselinname), 4 EU member states
- FOCUS Balkans data set: representatieve data sets (n=2,943) SPSS, 2010 (fruit, consumenten gedrag, zelf gerapporteerde voedselinname), 6 Balkan landen
- FOCUS Balkans data set Food Choice Motives: representative data set (n=2,943),SPSS, 2010, 6 Balkan landen.

In overleg met de consumenten onderzoekers zal er gekeken worden of er nog mogelijk andere, interessante data sets aanwezig zijn.

Dit project biedt een structuur waarin consumentengedragsonderzoekers via een aantal standaardstappen geholpen worden om data koppeling, verbanden tussen data sets en analyseproces te doorlopen. Tijdens de kick off bijeenkomst van het project wordt een begin gemaakt een procesoverzicht hoe onderzoekers op zoek gaan naar

verbanden tussen dataobjecten. In de analyse fase zal beschreven worden en hoe zij deze relaties kwantificeren, waarbij ook wordt ingegaan op de hulpmiddelen die dat proces, vooral indien de datasets groter, talrijker en dus onoverzichtelijker worden, kunnen helpen met inzicht krijgen in de dataset of sets. In de ontwikkelfase wordt oa. vanuit business intelligence het overzicht verder aangevuld met hulpmiddelen die passen bij de behoefte van de onderzoekers. In de testfase wordt ervaring opgedaan op gebied van data access, data science en valorisatie. Gedurende het gehele proces moet tevens gekeken worden welk valorisatie sporen interessant zijn, zoals 1) Ideeën genereren t.a.v. interessante verbanden, 2) Witte vlekken opsporen, 3) Tijd en geld besparen voor dataverzameling in vervolgonderzoek en/of 4) Kennis en inzichten genereren door m.b.v. algoritmes daadwerkelijk verbanden te onderzoeken. Al deze zaken geven ideeën voor de valorisatiestrategie op basis van de achterliggende vraag: kunnen we al "teasers / mockups" maken van mogelijke producten die we kunnen aanbieden voor verschillende klantgroepen?

Output

- Architectuur visie WUR specifiek CDP (volgens FAIR principes)
- Gedetailleerd ontwerp (samenhang, governance, business, informatie systeem en technologische infrastructuur)
- Werkend prototype CDP inclusief gebruikershandleiding
- Test – en evaluatie rapport, inclusief proof of principle en technische specificatie

Resultaat

In dit project worden de eerste bouwstenen uitgewerkt en getoetst (proof of principle) voor een toekomst bestendige data /gegevens infrastructuur als basis voor het uitrollen van een WUR CDP. De technische, organisatorische en business modellen resultaten uit het Europese project RICHFIELDS zullen als uitgangspunt gelden. Expertise op gebied van consumenten onderzoek, ICT en data science en business modellen en valorisatie wordt gezamenlijk opgepakt om goede integratie en een duurzame gegevens- en kennisinfrastructuur te waarborgen.

Betrokkenheid andere belanghebbenden

In het project zal samengewerkt worden tussen:

- WEcR en WFBR op gebied van interoperabel maken van consumenten data
- WR en WDCC op gebied van data science en infrastructuur (applied data science expertise);
- WR en FNH-RI om het Europese kennis m.b.t. technisch ontwerp en standaarden in het project te borgen;
- WR en WDCC in het uitwisselen van (technische) kennis m.b.t. CDP en WDCC;
- WR en ELIXIR in het uitwisselen van kennis m.b.t. het FAIR maken van (consumenten) data.

Kennisoverdracht

ELIXIR zal een twee daagse workshop organiseren en 2 dagen een advies ingehuurd worden. Dit project wordt niet afgesloten met een workshop voor consumenten onderzoekers, wel zal het proces, incl. de gebruikershandleiding onderdeel van een agenda kunnen worden in het reguliere overleg tussen WEcR en FBR consumenten onderzoekers. Het verder testen en uitrollen van het WUR CDP moet op een later tijdstip (2019) gebeuren.

Expertise

Karin Zimmermann: Overall projectleider; WEcR research manager FNH-RI**; Jos van de Puttelaar:** Onderzoeker consumer science, projectleider Smart Food Intake**; Ireen Raaijmakers:** Projectleider ENRICH**,** Onderzoeker consumer science**; Monique Vingerhoeds:** Onderzoeker consumer science IPR expert RICHFIELDS; **Garmt Dijksterhuis:** Onderzoeker Consumer Science**; Robbert Robbemond:** ICT specialist in relatie tot user needs; onderzoeker in RICHFIELDS en Prospect FNH-RI **Annette Breemer:** ICT specialist**; Jos Verstegen:** Business en governance modellen in relatie tot producten en valorisatie; onderzoeker in PPS Personalised Nutrition and Health en PPS DataFair NL. **Dominique van Wonderen:** WEcR Statisticus en data scientist

Economische en maatschappelijke relevantie

Deze use case geeft inzicht in de mogelijkheden van datasets, het koppelen van datasets en het gebruiken en analyseren van datasets om inzicht te krijgen in de valorisatiepotentie van een WUR Consumer Data Platform.

Rapportage

- Agenda en verslag Workshop Fair data
- Puntsgewijze notulen van afspraken van beslismomenten in het project
- Gebruikshandleiding
- Visie op vervolgstappen 2019

Risico analyse

De financiële verdeling ligt niet helemaal vast, afhankelijk van de vragen of expertise inzet in het project kan de financiële verdeling tussen categorieën van activiteiten aangepast worden

Project budget

Activiteit	Medewerker	Dagen	Kosten
A ELIXIR Workshop FAIR	Rob Van 't Hoofd / Celia van Gelderen/ ICT specialist ELIXIR; Max. 12/14 deelnemers; 2 dagen	6	6.000
B Data Bases	Jos van de Puttelaar en ?	17	17.000
C CDP	Robert Robbemond en Annette Breemer	17	17.000
D Data science en valorisatie	Jos Verstegen; Monique Vingerhoeds (WFBR); Dominique van Wonderen	13	15.000
E Management, ELIXIR advies en WDCC advies	Karin Zimmermann (4); Rob van 't Hoofd ELIXIR (2) Sjaak Wolfert (1)	7	7.000
Excl. BTW			62.000
Incl. BTW			75.000