Designprinzipien des IHR-15 RDF Data Stores

Projektgruppe „Interaktiver Haushaltsplanrechner Leipzig 2015“
http://www.leipzig-data.de/IHR-15

Version vom 25. September 2015

Inhaltsverzeichnis

Hintergrund

Das Projekt **Interaktiver Haushaltsplanrechner Leipzig 2015** ist ein wesentlicher Baustein des in enger Zusammenarbeit der Koordinierungsstelle für Bürgerbeteiligung der Stadt Leipzig („Leipzig weiter denken“), des Dezernats Finanzen der Stadt Leipzig und des Instituts für Öffentliche Finanzen und Public Management entwickelten Vorhabens **Nachhaltige Stadtfinanzen – Akzeptanzsteigerung der bürgerschaftlichen Beteiligung an der Haushaltsplanung**. Dieses Vorhaben wurde im Rahmen der Initiative „ZukunftsWerkStadt” im Zeitraum von Oktober 2014 bis August 2015 vom Bundesministerium für Bildung und Forschung (BMBF) durch Fördermittel unterstützt.

Als Teil der Strategie „Leipzig weiter denken 2.0“ war das Ziel des Vorhabens, die deliberativen Diskussions- und Beteiligungsstrukturen im Haushaltsplanungsprozess der Stadt Leipzig weiter zu stärken. Neben der im Rahmen von „Leipzig weiter denken“ bereits entwickelten repräsentativen Bürgerwerkstatt sollte deshalb eine noch intensivere bürgerschaftliche Einbindung ermöglicht werden. Für die Haushaltsentwurfsplanung bedeutet dies, das Handeln der Stadt noch transparenter zu gestalten und die Bürgerinnen und Bürger aktiver in Entscheidungsprozesse mit einzubeziehen. In diesem Zusammenhang wurde der von der Stadt Leipzig in den Jahren 2008 bis 2012 bereitgestellte, aber wenig genutzte „interaktive Haushaltsplan“ geprüft. Das Vorhaben wurde von der Koordinierungsstelle „Leipzig weiter denken“ beraten.

Aufbauend auf Ergebnissen aus Umfragen, Workshops und Good-Practiceanalysen wurde im Projektbaustein „Interaktiver Haushaltsplanrechner Leipzig 2015“ ein für Leipzig bedarfs- und zielgruppengerechtes Instrument erstellt. Anknüpfend an vorhandene Erfahrungen auch der Leipziger Agenda21 Gruppe bzw. des Forums Bürgerstadt Leipzig, deren Mitarbeiter die Entwicklung des „Interaktive Haushaltsplan“ unterstützt und begleitet haben, wurden Instrumente entwickelt, um haushaltsrelevante Informationen nutzergruppenfreundlich aufzubereiten und geeignete Partizipationsmöglichkeiten zu schaffen.

Projektpartner bei der Entwicklung seitens der Universität waren das Institut für Öffentliche Finanzen und Public Management (Prof. Lenk, Herr Redlich, Herr Glinka), das in der informationstechnischen Umsetzung durch das Institut für Informatik (Prof. Gräbe) bei der Anforderungsanalyse und der prototypischen technischen Realisierung eines neuen Online-Tools unterstützt wurde.

An der Realisierung des neuen Online-Tools arbeiteten die folgenden Studierenden mit:

> Wolfgang Amann, Janos Borst, Sarah Cujé, Christian Hoffmann, Dennis Kreußel, Fabian Niehoff, Tobias Wieprich, Tamara Winter, Kalle Willi Wollinger, Sebastian Zänker.

Die Arbeiten wurden weiterhin betreut von Prof. Gräbe und Konrad Höffner, Mitarbeiter am Lehrstuhl „Betriebliche Informationssysteme“, und Marius Brunnert als studentischer Tutor sowie durch Philipp Glinka und Matthias Redlich als projektverantwortliche Mitarbeiter am Institut für Öffentliche Finanzen und Public Management.

1 Allgemeines

Mit dem Übergang zum „Doppik“ haben sich 2012 auch in Leipzig die Grundsätze der Haushaltsführung geändert. Mit *Ergebnishaushalt* und *Finanzhaushalt* unterscheidet die „doppelte Buchführung“ zwischen der Darstellung von Einnahmen und Ausgaben (Ergebnishaushalt) sowie von Aufwendungen und Ergebnissen (Finanzhaushalt). Die Darstellung im veröffentlichten Haushalt der Stadt Leipzig (4 Bände pro Jahr, die als PDF online bei `http://leipzig.de` eingesehen werden können) nach verschiedenen Systematiken (Produktgruppen und Produkte, Finanzpositionen, Kostenstellen und Kostenarten) macht die Sache zusätzlich unübersichtlich.

Weiterhin sind in einem einzelnen Haushaltsplan nicht nur die Daten für das aktuelle Jahr berücksichtigt, sondern auch Prognosen für die Folgejahre ausgewiesen. Schließlich ist noch zu unterscheiden zwischen zu verschiedenen Zeitpunkten fixierten *Plandaten* und dem erst retrospektiv verfügbaren *Haushaltsergebnis*, so dass selbst für ein Jahr verschiedene Zahlenmaterialien im Umlauf sind.

> Bis zum Projektende konnten fundamentale Unstimmigkeiten in der Interpretation der durch das Dezernat für Finanzen der Stadt Leipzig zur Verfügung gestellten Haushaltsdaten nicht ausgeräumt werden, so dass die mit dem Prototyp ausgelieferten Daten **nur kursorisch die Funktionalitäten des Haushaltsplanrechners demonstrieren, nicht aber belastbare Informationen zur Haushaltssituation der Stadt Leipzig wiedergeben.**

Im folgenden Text wird ein Grundverständnis der kommunalen Haushaltssystematik vorausgesetzt.

2 Die Datenlage

Dem Projekt standen initial (seit Oktober 2014) seitens des Dezernats für Finanzen der Stadt Leipzig

(1) *Haushaltsplandaten zum Jahr 2014*

zur Verfügung, welche

- in einer Datei `Ratsinfo` die Haushaltssystematik in einer zusammenfassenden Darstellung bis zur Ebene 4 der *Bezüge*[1] sowohl des Ergebnishaushalts als auch des Finanzhaushalts abbildeten,
- in einer Datei `ErgHH_PB11_2014` beispielhaft detailliertere Daten für einen Produktbereich (PB 11) des Ergebnishaushalts sowie
- in einer Datei `FinHH` Daten zum Finanzhaushalt enthielten.

[1] So lautet die für Ergebnishaushalt und Finanzhaushalt gemeinsame Spaltenüberschrift in der Datenquelle.

In den Haushaltsplandaten waren Planansätze für die Jahre 2013 bis 2017 enthalten.

> Problematisch war von Anfang an die Interpretation der uns übergebenen Daten. Kumuliert über alle 1128 Posten des Produktbereichs 11 ergaben sich etwa folgende Summen für Einnahmen und Ausgaben im Planansatz 2014:
>
> | | `ErgHH_PB11_2014` | `Ratsinfo` |
> |---|---|---|
> | Einnahmen | 976.928.992 | 84.642.290 |
> | Ausgaben | − 938.157.157 | − 620.150 |
>
> Dagegen sind im veröffentlichten[2] Planansatz 2014 für den Produktbereich 11 ausgewiesen:
>
> - ordentliche Erträge: 25.148.860 Euro,
> - ordentliche Aufwendungen: − 99.757.851 Euro.

Das gesamte **Visualisierungskonzept** des Prototypen wurde auf der Basis der Struktur der Daten aus der Datei `Ratsinfo` zum Jahr 2014 erarbeitet. Dieses Visualisierungskonzept ist auch eng angelehnt an die Strukturierungskonzepte des bisherigen Haushaltsrechners und entspricht den Anforderungserhebungen aus den Workshops.

Aus diesen Daten ließ sich im Teil *Ergebnishaushalt* die Systematik der Produktbezeichnungen rekonstruieren (ein Bezug pro Zeile, jede Zeile ist mit Angaben zu Produktbereich, -gruppe und -untergruppe versehen). Diese 4-Ebenen-Systematik wird im folgenden als **Produktgraph** bezeichnet, die genauere Charakterisierung der einzelnen Knoten dieses Graphen als **Produktmodell**.

Am 27.07.2015 wurde in einem Gespräch mit dem Dezernat für Finanzen der Stadt Leipzig klar, dass die in der Systematik angegebenen Produkte nicht gleichwertig sind, sondern Produkte mit 10-stelligen Nummern die Rolle von *Schlüsselprodukten* spielen, zu denen auch zum großen Teil ausführlichere Produktsteckbriefe existieren, während andere Produkte als Unterprodukte anzusehen sind. Auf dieser Basis wurden in einem *Redesign des Produktmodells* allein die Schlüsselprodukte als vierte Ebene des Produktgraphen belassen, diese aber begrifflich von PSP-Elementen getrennt[3]. Das PSP-Element eines einzelnen Datensatzes der Primärdaten wird eindeutig einem Schlüsselprodukt zugeordnet, womit Einnahmen und Ausgaben bereits auf der Ebene der Schlüsselprodukte getrennt kumuliert werden können.

Eine solche Systematik konnte aus den Angaben im Teil *Finanzhaushalt* nicht rekonstruiert werden, da hier Bezüge in mehreren Zeilen vorkamen und in den meisten Fällen auch die Einordnung in den Produktgraphen fehlte.

Der Entwicklung unseres Prototyps wurden deshalb ausschließlich Plandaten des Ergebnishaushalts aus dem Jahr 2014 zu Grunde gelegt.

Im Mai 2015 wurde uns durch das Dezernat für Finanzen der Stadt Leipzig

(2) ein *Produktplan 2015 mit Spielraumangaben*

und im Juni 2015

[2] Haushaltsplan 2014 der Stadt Leipzig, Band 1, S. 531.

[3] Zu einer 10-stelligen Produktnummer gibt es nun sowohl ein PSP-Element als auch ein Schlüsselprodukt.

(3) eine *Übersicht mit 205 Produktsteckbriefen*

übergeben, aus denen sich weitere Informationen zur Systematik des Ergebnishaushalts entnehmen ließen:

- Im *Produktplan* sind in jeder Zeile entweder ein PSP-Element mit einer Bezeichnung `1.100.*` oder ein „Innenauftrag“ dargestellt. Die PSP-Nummern stimmen mit dem überein, was in der Ratsinfo-Darstellung im Bereich *Ergebnishaushalt* als *Bezug* bezeichnet ist.
- Unter Weglassung der Punkte ergeben sich mindestens 10-stellige PSP-Nummern, wobei sich die 10-stelligen PSP-Nummern zum überwiegenden Teil auch in den Produktsteckbriefen wiederfinden. Der Präfix 1100 weist darauf hin, dass es sich um ein Produkt aus dem Ergebnishaushalt handelt.
- Im Gespräch am 27.07.2015 wurde uns erstmals der Hintergrund erläutert – die 10-stelligen PSP sind *Schlüsselprodukte*, weitere PSP werden als *Unterprodukte* bezeichnet und haben eine Nummer, aus der sich die Nummer des zugehörigen Schlüsselprodukts als Präfix ergibt. Unterprodukte haben im Gegensatz zu Schlüsselprodukten keine ausführliche Beschreibung.
- Aus diesen PSP-Nummern lässt sich die Einordnung in den Produktgraphen unmittelbar ablesen.

Auf dieser Basis wurden die RDF-Graphen **NeuesProduktModell** (Produktmodell des Ergebnishaushalts), **NeuerProduktgraph** und **Produktsteckbriefe** im Rahmen eines Redesigns neu erstellt, in denen die Systematik des Ergebnishaushalts für unseren Kontext umfassend dargestellt ist. Diese Darstellung ist mit dem Produktplan (S. 686 ff. Haushaltsplan, Band 1) abgeglichen.

Im Mai 2015 wurden uns seitens des Dezernats für Finanzen der Stadt Leipzig *detaillierte Daten zum Haushaltsplan 2015/16* übergeben, die jedoch einer anderen Systematik folgten als die von uns verwendeten Plandaten 2014 und sich damit nicht in das Visualisierungskonzept entsprechend unserer Anforderungserhebung einordnen ließ.

Unsere Bemühungen zur Aufbereitung dieser Daten können hier außer Betracht bleiben, da die Datenbasis am 27.07.2015 noch einmal aktualisiert wurde. Das Dezernat für Finanzen der Stadt Leipzig übergab uns an diesem Tag neu zusammengestellte Daten zu den Haushalten 2015 und 2016 einschließlich kumulierter Übersichten `Ratsinfo` für diese beiden Jahre (allerdings mit fehlerhafter Zeichensatz-Kodierung).

Die dort enthaltenen **Planwertdaten für die Ergebnishaushalte der Jahre 2014 bis 2019** bildeten die Basis für weitere Datentransformationen, mit denen die Datenbasis des Prototypen im letzten Projektmonat August 2015 komplett erneuert und im Rahmen eines allgemeinen Redesigns in die finale Version integriert wurde.

> Auch in diesen Daten gibt es erhebliche Differenzen. Kumuliert über alle 18.848 Posten des aus der Primärquelle `ErgHH_2015` extrahierten detaillierten Ergebnishaushalts bzw. über alle 1.158 Posten des in der Datei `Ratsinfo_2015` dargestellten Ergebnishaushalts ergaben sich folgende Summen für Einnahmen und Ausgaben im Planansatz 2015:

	ErgHH_2015	Ratsinfo
Einnahmen	– 3.399.306.619,31	– 1.487.889.183,26
Ausgaben	3.384.184.302,47	1.472.766.866,42

Es fällt auch auf, dass gegenüber den digitalen Haushaltsdaten von 2014 sowie der veröffentlichten pdf-Version des Doppelhaushalts die Vorzeichen von Einnahmen und Ausgaben vertauscht sind. Diese und Fragen zu weiteren Inkonsistenzen in den Daten konnten während der Projektlaufzeit nicht geklärt werden.

Da in der Datei `Ratsinfo_2015` eine detaillierte Gegenüberstellung von Einnahmen und Ausgaben nur für 2015 gegeben wird, für Vergleichsjahre aber nur eine konsumierte Sicht dargestellt ist, im detaillierten Ergebnishaushalt dagegen in den Planansätzen für die Jahre 2014 bis 2019 Einnahmen und Ausgaben gegenübergestellt werden, wurde das Redesign auf der Basis der detaillierten Daten aus den Primärquellen `ErgHH_2015` und `ErgHH_2016` ausgeführt, um die Möglichkeiten der visuellen und tabellarischen Gegenüberstellung von Haushaltsdaten verschiedener Jahre auszuloten.

Im Gespräch am 27.07.2015 wurde auch besprochen, dass für den Prototyp die Einnahmen und Ausgaben auf der Ebene der Schlüsselprodukte aggregiert werden sollen, da nur für diese Produktsteckbriefe vorliegen und damit eine ausreichende Informationsbasis für Erläuterungen zur Verfügung steht, die im Rahmen der Anforderungserhebungen als hoch priorisiert eingestuft wurden.

Für eine systematische Transformation der Datenbasis der **Finanzhaushalte** sind ähnlich viele Fragen offen, weshalb diese Problematik im Projekt nicht in Angriff genommen wurde.

3 RDF Data Cube – Konzeptionelle Vorbemerkungen

Im weiteren Text wird davon ausgegangen, dass die grundlegenden Konzepte des RDF Data Cube Ansatzes [2] bekannt sind. `qb:` steht wie üblich für den RDF Cube Namensraum.

Die Datenbasis des Prototyps ist in einzelne *RDF Data Cubes* als Datasets aufgeteilt, in denen jeweils alle Datensätze (Observations) aus einer Quelle und zu einem Jahr zusammen mit dem Verweis auf die formale Beschreibung der Struktur dieser Observations zusammengefasst sind.

Ein **Datenset** besteht aus

- einer Serie von RDF Data Cubes mit den Daten für jeweils ein Jahr,
- einem RDF Graphen *Config*, in dem beschrieben ist, welcher RDF Data Cube welchem Jahr zugeordnet ist, sowie
- RDF Graphen mit dem Produktgraphen, dem Produktmodell, den Produktbeschreibungen und der Dataset-Beschreibung der RDF Data Cubes.

3.1 RDF Data Cubes und deren Beschreibung

Ein RDF Data Cube enthält *ein DataSet* (eine Instanz des RDF-Typs `qb:DataSet`) sowie gleichartig strukturierte *Observations* (Instanzen vom RDF-Typ `qb:Observation`), die alle einer gemeinsamen *Modellstruktur* für die Datenerfassung folgen, deren Beschreibung sowohl syntaktisch als auch semantisch als eine *DataStructureDefinition* über das Prädikat

`qb:structure` des DataSets referenziert ist. Im folgenden Code-Beispiel ist der Zusammenhang zwischen einer Observation `ihrdata:EH_15G_Plan14-Bezug1100111101` und dem Dataset `ihrds:EH_15G_Plan14` aus einem unserer Cubes prototypisch dargestellt.

```
ihrds:EH_15G_Plan14 a qb:DataSet ;
  rdfs:label "Haushaltsdaten 2015 der Stadt Leipzig ..."@de ;
  rdfs:comment "..."@de ;
  dct:source "Dezernat für Finanzen der Stadt Leipzig" ;
  qb:structure ihr:DSDShort .

ihrdata:EH_15G_Plan14-Bezug1100111101 a qb:Observation ; ...
  qb:dataSet ihrds:EH_15G_Plan14 .
```

Um statistische Vergleichbarkeit zu erreichen, folgt eine zusammengehörende Serie von solchen *DataSets* derselben Modellstruktur, also derselben *DataStructureDefinition*, die im Fall des Prototyps die URI `ihr:DSDShort` hat und in der Datei `HaushaltLeipzigCube.ttl` definiert ist, die wir uns nun genauer anschauen wollen.

Eine *DataStructureDefinition* beschreibt ein RDF-Schema, das seinerseits das RDF-Schema der Observations beschreibt. Die Beschreibung dieser Beschreibung (Metamodellebene) setzt dabei konsequent auf den (Meta)-Konzepten von RDF-Modellen [1] wie `rdfs:subClassOf` und `rdfs:subPropertyOf` auf. Eine *DataStructureDefinition* besteht aus einer Menge von Instanzen vom RDF-Typ `qb:ComponentSpecification`, die der *DataStructureDefinition* über das Prädikat `qb:component` zugeordnet sind und mit denen die in den Observations verwendeten *Prädikate* genauer beschrieben werden. Hierbei wird ein besonderes syntaktisches Moment von RDF ausgenutzt – Prädikate einer Ebene können auf einer Metaebene selbst *Subjekt* von Beschreibungen sein.

3.2 Optional: ComponentSpecification und ComponentProperties

Um solche Prädikate zu verschiedenen Prädikatklassen zusammenzufassen, wird ein weiteres Indirektionsprinzip angewendet, mit dem Vererbungsstrukturen in RDF modelliert werden: Über das RDF-Oberprädikat `qb:componentProperty` ist einer *ComponentSpecification* eine Instanz vom RDF-Obertyp `qb:ComponentProperty` zugeordnet, von denen es verschiedene Unterprädikate mit zugehörigen Subtypen als Wertebereich (Range) gibt:

- `qb:dimension` mit Wertebereich `qb:DimensionProperty`,
- `qb:attribute` mit Wertebereich `qb:AttributeProperty`,
- `qb:measure` mit Wertebereich `qb:MeasureProperty`.

Im folgenden Beispiel ist in der DataStructureDefinition `ihr:DSDShort` eine ComponentSpecification `ihr:jahrComponent` definiert, die ihrerseits eine spezielle DimensionProperty `ihr:jahr` definiert, die den Wertebereich `xsd:gYear` hat und als Prädikat in Observations verwendet werden kann.

```
ihr:DSDShort a qb:DataStructureDefinition ;
  rdfs:label "Data Structure Definition für den Leipziger Haushalt"@de ;
  qb:component ihr:jahrComponent .
```

```
ihr:jahrComponent a qb:ComponentSpecification ;
  rdfs:label "Jahr-Component"@de ;
  qb:dimension ihr:jahr .

ihr:jahr a rdf:Property, qb:DimensionProperty ;
  rdfs:label "Jahr"@de ;
  rdfs:range xsd:gYear .
```

Jeder *ComponentSpecification* ist auf diese Weise eine *ComponentProperty* zugeordnet, die zu einer der drei definierten Unterklassen gehören kann, was daran zu erkennen ist, welches der Prädikate `qb:dimension`, `qb:attribute` oder `qb:measure` in der Definition verwendet wird.

Diese auf den ersten Blick etwas umständliche Indirektion macht sich erforderlich, weil RDF nur 3-Wort-Sätze kennt. Explizite URIs für *ComponentSpecifications* können durch „blank nodes“ umgangen werden, allerdings folgen wir der Empfehlung, „blank nodes“ aus anderen syntaktischen Gründen konsequent zu vermeiden.

3.3 Component Properties

Wichtig für das weitere Verständnis der Struktur der Observations sind allein die *ComponentProperties* in ihren drei Typ-Ausprägungen als *DimensionProperties*, *AttributeProperties* und *MeasureProperties*. Jede *ComponentProperty* ist insbesondere eine `rdf:Property` mit den Prädikaten `rdfs:label`, `rdfs:comment` und `rdfs:range`.

Als `rdfs:range` können standardmäßige RDF-Typbezeichner-Klassen verwendet werden. Oft müssen aber für spezielle Modelle weitere Klassen definiert werden, insbesondere als Wertebereich für *DimensionProperties*. Für derartige Definitionen wurde das Konzept des `qb:measureType` entwickelt, um einen generischen Rahmen für entsprechende Typklassen zu entwickeln. Dieses Konzept kommt bei uns derzeit nicht zum Einsatz.

Für die eingeführten Properties sind neben syntaktischen auch semantische Aspekte zu definieren. Dies erfolgt mit dem Prädikat `qb:concept`, über welches jeder ComponentProperty (und damit auch jeder der drei möglichen UnterProperties) eine formal definierte Semantik aus geeigneten Konzeptwerken zugeordnet werden kann. Im IHR-Modell wird hierfür konsequent das im Statistikbereich inzwischen etablierte SDMX-Vokabular [3] referenziert.

4 Zur Struktur unserer Daten

4.1 Die Struktur unserer RDF Data Cubes

Jeder unserer RDF Data Cubes enthält eine Menge von Observations (jeweils aus einer spezifizierten Primärquelle zu einem Jahr) zusammen mit der zugehörigen DataSet-Definition und einer Beschreibung des RDF-Graphen als Instanz von `owl:Ontology`. Dazu sind zwei Namensraum-Präfixe

- `ihrdata: <http://haushaltsrechner.leipzig.de/Data/Observation/>` und
- `ihrds: <http://haushaltsrechner.leipzig.de/Data/Dataset/>`

definiert, die als Präfixe für URIs der Observations (`ihrdata`) bzw. des Datasets (`ihrds`) dienen.

Die Bezeichnungen des RDF Graphen, des dort enthaltenen Datasets und der einzelnen Observations folgen einem einheitlichen Muster, das von einer *Kennung* des Cubes ausgeht. Für die Kennung `EH_15G_Plan14` (Ergebnishaushalt Plandaten 2014, extrahiert aus den Haushaltsdaten 2015/16, die uns Ende Juli 2015 übergeben wurden) lauten die URIs wie folgt:

- `http://haushaltsrechner.leipzig.de/Data/EH_15G_Plan14/` ist die URI des RDF Graphen, der auch als Turtle-Datei `EH_15G_Plan14.ttl` im git-Repo gespeichert ist;
- `ihrds:EH_15G_Plan14` ist die URI des Datasets und
- `ihrdata:EH_15G_Plan14-Bezug1100111101` ist die URI einer Observation, die sich aus der gewählten Kennung und der Produktnummer – in diesem Fall eines Schlüsselprodukts – zusammensetzt, auf die sich die jeweilige Observation bezieht.

  Da es in jedem Dataset zu jedem Produkt nur einen Datensatz gibt, wird auf diese Weise die eindeutige Referenzierbarkeit der Observations auch über die Grenzen des jeweiligen Datasets hinaus gewährleistet.

Unter der URI des RDF Graphen ist die genaue Primärdatenquelle auch noch einmal als `rdfs:comment` angegeben.

Eine *Observation* hat typischerweise folgende Struktur:

```
ihrdata:EH_15G_Plan14-Bezug1100111101
  ihr:relatesTo ihr:Bezug1100111101 ;
  ihr:jahr "2014";
  ihr:kategorie ihr:Ergebnishaushalt ;
  ihr:ein "2341177.24"; ihr:aus "-153900";
  ihr:waehrung dbpedia:Euro ;
  qb:dataSet ihrds:EH_15G_Plan14 ;
  a qb:Observation .
```

Die einzelnen Prädikate sind in der *DatasetDefinition* `ihr:DSDShort` definiert, die im DataSet `ihrds:EH_15G_Plan14` referenziert wird:

- `ihr:relatesTo` – eine *DimensionProperty*, die auf die Produktnummer verweist, auf welche sich die *Measure* bezieht,
- `ihr:jahr` – eine *DimensionProperty*, die auf das Jahr verweist, auf welches sich die *Measure* bezieht[4],
- `ihr:kategorie` – eine *DimensionProperty*, die auf den Haushalttyp verweist, auf welchen sich die *Measure* bezieht,
- `ihr:ein`, `ihr:aus` – zwei *MeasureProperties*, unter denen die Einnahmen und Ausgaben zu diesem Sachverhalt zahlenmäßig als `xsd:decimal` erfasst sind,
- `ihr:waehrung` – eine *AttributeProperty*, die angibt, auf welche Einheit sich die Maßzahlen beziehen. Wert dieser Property ist in allen Fällen `dbpedia:Euro`, eine Referenz[5] auf die Definition der Währungseinheit „Euro“ in der weit verbreiteten Dbpedia-Ontologie[6].

[4] Der Range dieser Property ist `xsd:gYear`, allerdings wird auf die Range-Angabe als Datentyp in den einzelnen Cubes verzichtet, da diese Angaben beim Ontowiki-Import aktuell falsch behandelt werden.

[5] `http://dbpedia.org/resource/Euro`

[6] `http://wiki.dbpedia.org/use-cases/multi-domain-ontology`

In jedem RDF Data Cube sind zusammen mit einer Observation zu einem Knoten im Produktbaum auch zu jedem Vorgängerknoten bis hinauf zum Wurzelknoten `ihr:`*Kennung*`-Top` je eine Observation enthalten. Einnahmen und Ausgaben in einer Observation zu einem Knoten im Inneren des Produktbaums ergeben sich jeweils als Summe der Einnahmen bzw. Ausgaben aller Kindknoten zu diesem Knoten. **Die Einnahmen und Ausgaben sind also auf allen Ebenen des Produktbaums aus den Angaben in den Blättern kumuliert vorberechnet.**

4.2 Das Produktmodell

Wie bereits im Abschnitt „Die Datenlage“ erläutert, wird unser Produktmodell der Systematik des Ergebnishaushalts in einem **Produktgraphen** beschrieben, in welchem die genauere Charakterisierung der einzelnen Knoten dieses Graphen als **Produktmodell** formalisiert ist. Beide liegen als RDF-Graphen im Turtle-Format vor, der Produktgraph in der Datei `NeuerProduktgraph.ttl` und das Produktmodell in der Datei `NeuesProduktModell.ttl`.

Das Produktmodell umfasst die vier Ebenen Produktbereiche, Produktgruppen, Produktuntergruppen und Schlüsselprodukte. Für jede dieser vier Ebenen ist ein RDF-Typ definiert. Die URIs der Instanzen des jeweiligen Typs werden aus einem typspezifischen Präfix und einer ID in der Form `ihr:<Präfix><ID>` gebildet. Die entsprechenden Informationen wurden für den Ergebnishaushalt aus verschiedenen Quellen extrahiert und mit dem Produktplan (S. 686 ff. Haushaltsplan 2015/16, Band 1) abgeglichen.

Die Details sind in der folgenden Tabelle zusammengestellt.

Anzahl	Ebene	Präfix	RDF-Typ
32	Produktbereiche	PrBer	`ihr:PrBer`
92	Produktgruppen	PrGr	`ihr:PrGr`
127	Produktuntergruppen	PrUGr	`ihr:PrUGr`
282	Schlüsselprodukte	Bezug	`ihr:Schluesselprodukt`

Hinzu kommt der Wurzelknoten `ihr:Top` von der RDF-Oberklasse `ihr:Node`, zu der alle Knoten des Produktbaums gehören. Die Beziehungen zwischen den Elementen benachbarter Ebenen sind als Kanten des (gerichteten) Produktgraphen durch das Prädikat `ihr:hasChild` beschrieben.

Im Redesign wurden die Schlüsselprodukte konsequent von den **PSP-Elementen** getrennt und für letztere ein weiterer RDF-Typ `ihr:PSP-Element` eingeführt, der über ein Prädikat `ihr:belongsTo` einem Schlüsselprodukt zugeordnet werden kann. Damit besteht in Zukunft die Möglichkeit, die Haushaltsdaten nicht nur auf der Ebene der Schlüsselprodukte zu kumulieren, sondern auch noch den Bezug zur Ebene der PSP-Elemente genauer darzustellen. Dies ist im aktuellen Prototyp allerdings nicht umgesetzt, sondern nur als RDF Graph `PSP-Elemente.ttl` im Turtle-Format als Erweiterung des Produktmodells in den Daten enthalten.

Zu den drei Ebenen Produktbereich, Produktgruppe und Produktuntergruppe liegen typischerweise Informationen in folgender Form vor:

```
ihr:PrUGr7520 a ihr:PrUGr ;
  rdfs:label "Schadensereignisse Bau- und Grundstücksordnung"@de ;
```

```
  ihr:BezeichnungStadt "Schadensereignisse Bau- und Grundstücksordnung" ;
  ihr:hatStadtId "7520" .
```

Hierbei sind den einzelnen Prädikaten folgende Informationen als Werte zugeordnet:

- `ihr:PrUGr7520` die URI des Produkts, die sich in diesem Fall einer Produktuntergruppe aus dem Präfix `ihr:PrUGr` und der ID der Untergruppe in der Stadtsystematik zusammensetzt,
- `ihr:PrUGr` der RDF-Typ des Produkts,
- `rdfs:label` eine aus den Stadtdaten entnommene und ggf. korrigierte oder aus anderen Quellen konsolidierte Bezeichnung des Produkts,
- `ihr:BezeichnungStadt` die aus den Stadtdaten übernommene genaue Bezeichnung mit allen Punkten, Abkürzungen und Sinnentstellungen, die sich aus den Größenbeschränkungen der Felder in der Primärquelle ergeben. Diese Bezeichnung dient bei der Suche in Primärdaten als Fremdschlüssel, um Bezüge zwischen verschiedenen Datenquellen aufzudecken.
- `ihr:hatStadtId` die Referenznummer des Produkts in den Stadtdaten.

Letzteres ist für Produktbereiche, Produktgruppen und Produktuntergruppen identisch mit der aus der Stadtsystematik inferierten ID, für Schlüsselelemente und PSP-Elemente wird die ID durch Entfernen der Punkte auf einen reinen Zahlenwert reduziert – in diesem Fall ist der Wert des Prädikats `ihr:hatStadtId` die ursprüngliche Referenz mit Punkten.

Schlüsselelemente enthalten weitere Informationen wie in folgendem Beispiel:

```
ihr:Bezug1100111102 a ihr:Schluesselprodukt ;
  rdfs:label "Leitungshilfe und Unterstützung"@de ;
  ihr:BezeichnungStadt "Leitungshilfe und Unterstützung" ;
  ihr:hatStadtId "1.100.11.1.1.02" ;
  ihr:hatGrad "2" ;
  ihr:zumAmt ihr:Amt011 .
```

wobei

- `ihr:hatGrad` den finanziellen Gestaltungsspielraum der Kommune bei diesem Schlüsselprodukt angibt und
- `ihr:zumAmt` eine Referenz auf eine URI des Amts entsprechend der Ämterübersicht (S. 634 ff. Haushaltsplan, Band 1) ist, welchem dieses Schlüsselprodukt zugeordnet ist.

Der **Gestaltungsspielraum** g ist eine rationale Zahl im Bereich $1 \leq g \leq 3$, wobei

1 für nicht beeinflussbar, 2 für bedingt beeinflussbar und 3 für stark beeinflussbar

steht. Da g in der Primärquelle `Produktplan_2015_mit_Spielraumangaben` teilweise nicht direkt den Schlüsselprodukten zugeordnet war, sondern nur verschieden ausgezeichneten Unterprodukten, sind die Werte g für einige Schlüsselprodukte interpoliert und dann auch nicht

mehr ganzzahlig. Da die Gestaltungsspielräume für einzelne Knoten höherer Ebenen im Produktgraphen sowieso auf eine noch zu bestimmende Weise aus den Gestaltungsspielräumen der Kindknoten zu berechnen sind, ist diese Setzung naheliegend. Eine solche Interpolation wurde allerdings im Projekt nicht durchgeführt, womit Spielraumangaben allein für die Ebene der Schlüsselprodukte vorliegen und in den jeweiligen Produktbeschreibungen mit angegeben werden.

4.3 Der Produktgraph

Zur Aufstellung einer ersten Version des Produktgraphen wurden alle Zeilen mit 10-stelligen PSP-Nummern (Schlüsselprodukte) aus der Datei `HH2014/EH_14Ratsinfo.csv` extrahiert, aus ihnen die Verbindungen von Produktbereich, Produktgruppe, Produktuntergruppe und Schlüsselprodukt entnommen und mit dem Prädikat `ihr:hasChild` beschrieben.

Ein typischer Eintrag sieht wie folgt aus

```
ihr:PrBer42 ihr:hasChild ihr:PrGr421, ihr:PrGr424 .
```

und ist mit Verweis auf die obigen Ausführungen zum Produktmodell weitgehend selbsterklärend. Dieser RDF Graph dient dazu, die Beziehungen zwischen den verschiedenen Einheiten auf der Ebene der URIs darzustellen. Er wurde im weiteren Projektverlauf mehrfach mit anderen Quellen abgeglichen und auf diese Weise konsolidiert.

5 Die Datentransformation der Primärdaten

Dem Projekt wurden folgende Primärdaten in Form von Exceldateien zur Verfügung gestellt, die aus dem SAP-System des Dezernats für Finanzen der Stadt Leipzig extrahiert wurden:

- Zu Projektbeginn im Oktober 2014 Plandaten zum Haushaltsansatz 2014, siehe Verzeichnis[7] `Primaerdaten/HH2014`,
- im Mai 2015 Plandaten zum Haushaltsansatz 2015, siehe Verzeichnis `Primaerdaten/HH2015-1` und
- Ende Juli 2015 Plandaten zu den Haushaltsansätzen 2015 und 2016, siehe Verzeichnis `Primaerdaten/HH2015-2`.

Diese Excel-Dateien wurden in der Regel nach folgendem Schema aufbereitet:

- Umwandlung der Excel-Datei in eine csv-Datei,
- Extraktion relevanter Datensätze für jeweils ein Jahr über das Kommando `grep`, das Zeilen nach vorgegebenem Muster (etwa „Planansatz2017“) aus der csv-Datei auswählt, und
- Transformation dieser Daten in RDF-Graphen mit verschiedenen Perl-Skripten aus dem Verzeichnis `Daten/Werkbank`, siehe Abschnitt 5.4.

[7] Die Primärdaten sind nicht Teil des finalen Releases.

5.1 Verzeichnis `Primaerdaten/HH2014`

Dieses Verzeichnis enthält die zum Projektbeginn im Oktober 2014 übergebenen Plandaten zum Haushaltsansatz 2014.

- `ErgHH_PB11_2014`

  Detaillierte Planwertdaten zum Produktbereich 11 des Ergebnishaushalts für die Jahre 2013 bis 2017.

- `FinHH`

  Datensätze zum Finanzhaushalt für die Jahre 2013 bis 2017.

- `Ratsinfo`

  Datensätze mit Jahresvergleichen (1128 Datensätze zum Ergebnishaushalt, 3947 Datensätze zum Finanzhaushalt)

Die Datensätze aus `Ratsinfo` wurden zur weiteren Verarbeitung in die Dateien

`EH_14Ratsinfo.csv` (Datensätze zum Ergebnishaushalt) und
`FH_14Ratsinfo.csv` (Datensätze zum Finanzhaushalt)

separiert.

5.2 Verzeichnis `Primaerdaten/HH2015-1`

Dieses Verzeichnis enthält die im Mai 2015 übergebenen Plandaten zum Haushaltsansatz 2015

- `ErgHH_1`

  Detaillierte Planwertdaten zum Ergebnishaushalt für die Jahre 2014 bis 2018 sowie Rechnungsendbeträge 2013.

- `FinHH_lfd_Verwaltungsteaetigkeit`

  Datensätze zum Finanzhaushalt aus laufender Verwaltungstätigkeit für die Jahre 2014 bis 2018 sowie Rechnungsendbeträge 2013.

- `invest_F8`

Transformationen dieser Datensätze wurden mit Blick auf die später als `HH2015-2` zur Verfügung gestellten aktuelleren Daten ähnlicher Struktur nicht weiterverwendet und sind deshalb auch nicht mehr im git-Repo enthalten.

5.3 Verzeichnis `Primaerdaten/HH2015-2`

Dieses Verzeichnis enthält die Ende Juli 2015 übergebenen Plandaten zum Haushaltsansatz 2015 und 2016

- `ErgHH_2015` und `ErgHH_2016`

  Detaillierte Planwertdaten zum Ergebnishaushalt für die Jahre 2014 bis 2019 sowie Rechnungsendbeträge 2013 und 2014.

- `FinHH_lfd.Vw_2015` und `FinHH_lfd.Vw_2016`

  Datensätze zum Finanzhaushalt aus laufender Verwaltungstätigkeit für die Jahre 2014 bis 2019 sowie Rechnungsendbeträge 2013 und 2014.

- `invest_2015` und `invest_2016`

  Datensätze zum Investitionshaushalt für die Jahre 2014 bis 2019 sowie Rechnungsergebnisse 2013 und 2014.

- `Ratsinfo_2015` und `Ratsinfo_2016`

  Datensätze mit Jahresvergleichen zum Ergebnishaushalt und zum Finanzhaushalt

Die Datensätze aus `ErgHH_2015` und `ErgHH_2016` wurden zur weiteren Verarbeitung in die Dateien

$$\texttt{EH\_15G\_Plan<j>.csv} \text{ mit } j \in \{14, 15, 16, 17, 18, 19\}$$

separiert. Die Plandaten zu einem Jahr, das in beiden Dateien erfasst ist, waren identisch, so dass hier keine weitere Unterscheidung erforderlich ist.

Die Datensätze zum Ergebnishaushalt aus `Ratsinfo_2015` und `Ratsinfo_2016` wurden zur weiteren Verarbeitung in die Dateien `EH_15Ratsinfo.csv` und `EH_16Ratsinfo.csv` separiert. Hierbei ist zu beachten, dass in den Ratsinfo-Dateien nur für die Daten des jeweiligen Jahrs nach Einnahmen und Ausgaben unterschieden wird, die Daten der Folgejahre dagegen nur in konsumierter Sicht vorliegen, wo Einnahmen und Ausgaben zu einer Zahl zusammengezogen sind. Deshalb können aus einer solchen Datei auch nur Einnahmen und Ausgaben für das jeweils aktuelle Jahr entnommen werden.

5.4 Transformation der csv-Jahresdateien

Mit Blick auf die unbefriedigende Datenlage, immer wieder geänderte Formate und ungeklärte Inkonsistenzen in den Primärdaten konnte im Rahmen des Projekts kein konsolidierter Transformationsprozess spezifiziert werden. Die im finalen Release ausgelieferten Transformationswerkzeuge in der Skriptsprache „Perl“

- `parseErgHH.pl` – Parser-Routinen für csv-Dateien zum Ergebnishaushalt sowie
- `parseFinHH.pl` – Parser-Routinen für csv-Dateien zum Finanzhaushalt

enthalten die gesammelten Best Practise Erfahrungen zu den ausgeführten Transformation und zeigen dem Profi, wie Transformationen innerhalb des Projekts ausgeführt wurden. Mit Blick auf fehlende Spezifikationen sind diese Werkzeuge zu keinem speziellen Zweck einsetzbar, sondern werden nur „as is“ ausgeliefert.

In jedem Fall wird davon ausgegangen, dass in einem Vorprozess aus den ins csv-Format verwandelten Primärdaten die Datensätze zu den einzelnen Jahren in separate csv-Dateien extrahiert wurden, die dann mit einem Perl-Skript weiterverarbeitet werden.

Im Transformationsprozess für die Plandaten aus dem Ergebnishaushalt des jeweiligen Jahres werden die Zahlen aus den Spalten „Einnahmen“ und „Ausgaben“ sowie die PSP-Nummer übernommen und wie folgt weiterverarbeitet:

- Aus der PSP-Nummer werden die zugehörigen URIs von „Schlüsselprodukt“, „Produktuntergruppe“, „Produktgruppe“ und „Produktbereich“ generiert.
- Die Zahlen aus den Spalten „Einnahmen“ und „Ausgaben“ werden in das internationale Zahlenformat transformiert.
- Diese Zahlen werden auf einem Hash mit den jeweiligen URIs als Schlüssel auf jeder Ebene des Produktbaums bis hin zum Wurzelknoten aggregiert.
- Der Hash wird ausgewertet, eine Observation pro Schlüssel erzeugt und als RDF Cube abgespeichert.

Der RDF Cube hat denselben Namen wie die Quelle, aus der er erzeugt wurde.

Für den aktuellen Haushalt wird mit der Cube-Serie `EH_15G_Plan*` gearbeitet, als zweite Serie steht `EH_*Ratsinfo` zur Verfügung. Die Datenbasis des Prototyps kann in der Datei `Config.ttl` konfiguriert werden.

Literatur

[1] RDF Schema 1.1. W3C Recommendation 25 February 2014. `http://www.w3.org/TR/rdf-schema/`.

[2] The RDF Data Cube Vocabulary. W3C Recommendation 16 January 2014. `http://www.w3.org/TR/vocab-data-cube/`.

[3] The Statistical Data and Metadata Exchange (SDMX) Initiative. `http://www.sdmx.org`.