# Aufbau eines Agrardatenzentrums in der Bundesanstalt für Landwirtschaft und Ernährung (BLE)

Carsten Schwarz, Christof Ansorge, Andreas Maul, Jan Pohlmann

Wissensmanagement und Planungsgrundlagen
Bundesanstalt für Landwirtschaft und Ernährung (BLE)
Deichmanns Aue 29
53179 Bonn
carsten.schwarz@ble.de

**Abstract:** Der Datawarehouse-Prozess dient zur Konsolidierung der Datenhaltungssysteme. Die Daten werden von verschiedenen Datenquellen bereitgestellt und im ETL-Prozess (Extraktion, Transformation, Laden) in das Datawarehouse geladen. Die Erstellung eines Datawarehouse basiert auf zwei Grundideen: zum einen der Integration von Daten aus verteilten und unterschiedlich strukturierten Datenbeständen und zum anderen der Separation von Daten, die für das operative Geschäft genutzt werden. Somit können Daten nicht nur als Zeitreihen, sondern auch im Zusammenhang verschiedener inhaltlicher Dimensionen vorgehalten werden.

## 1 Zielsetzung

Jede statistische Datensammlung muss prinzipiell mindestens drei Fragen beantworten:

1. Werden die benötigten Daten gesammelt?
2. Haben die Daten eine hinreichende Qualität?
3. Stehen die Daten rechtzeitig zur Verfügung?

Je nach Fragestellung sind unterschiedliche Daten von Bedeutung. Fragestellungen entwickeln sich aus einer Situation heraus oder stehen im Zusammenhang mit Aufträgen, wodurch sich immer ein gewisser Zeitvorlauf ergibt, der für die Sammlung genutzt werden kann. Entscheidend für die o.g. Fragen ist somit der zeitige Beginn der Sammlung, möglichst während sich die Fragestellung entwickelt. Für die Qualität der Daten ist dabei meist wichtiger, die genaue Entstehung zu kennen und zu dokumentieren, um die Aussagekraft zu kennen, anstatt wie bisher auf eine bestimmte Quelle vertrauen zu müssen.

Vor diesem Hintergrund ist es das Ziel, über Behördengrenzen hinweg eine Zusammenarbeit zu initiieren, mit der zentral ein Grunddatenbestand

- entsprechend dem *Bedarf* der Nutzer bereit gestellt werden kann,
- wobei die *Qualität* wissenschaftlichen Standards entsprechend beschrieben wird und
- die Daten rechtzeitig zur *Verfügung* gestellt werden.

In diesem Datenzentrum, welches nicht nur Primärdaten aus der eigenen Tätigkeit der Datensammlung aufnimmt, sollen zur Ergänzung des Kontextes auch Daten anderer Sammlungen in einer einheitlichen Struktur aufgenommen werden, um

> Daten aus unterschiedlichen fachlichen Bereichen miteinander verschneiden zu können.

Indem über Behördengrenzen hinweg kooperiert wird, werden wiederholte, aus unterschiedlichem Anlass sich ergebende Sammlungen vermieden, womit

> eine *rationelle Verwendung der Arbeitskapazitäten* unterstützt wird.

## 2 Herangehensweise

Während das Datenzentrum ausschließlich für die Informationsbereitstellung genutzt werden soll, kann die Struktur unabhängig von operativen Verfahren gestaltet werden. Beim Datenpool werden die historischen Daten aus der operativen Vorgangsbearbeitung bereitgestellt. Aus der Tatsache heraus, dass in dem Datenzentrum der Zukunft noch zu definierende Sachverhalte und fachliche Bereiche aufzunehmen sind, die in einer einheitlichen Struktur mit den bisher gesammelten Daten stehen müssen (Verschneiden unterschiedlicher Datenbereiche), ist besondere Sorgfalt auf die Definition der Datenstruktur zu legen. Ausgehend von einer exemplarisch gefundenen Struktur aus bereits vorhandenen Datenbeständen, ist zu erwarten, dass diese um neue Bereiche ergänzt werden kann und so auch zukünftig in jeder Hinsicht gerecht wird. Der Erkenntnisgewinn, der sich aus der Kombination unterschiedlichster Datenbereiche ergibt, bleibt so lange verschlossen, wie eine technische Verfügbarkeit in einer gleichartigen Struktur nicht gewährleistet werden kann.

Auf der Grundlage der bisherigen Erfahrung ist somit eine abstrakt zu definierende Datenstruktur zu finden, die mit hoher Sicherheit verschiedensten fachlichen Bereichen mit einer einheitlichen Struktur gerecht wird. Aus diesem Grund wird dem eigentlichen Projekt die Entwicklung eines Prototyps eines Datawarehouses vorangestellt, in dem die Grundstruktur der Datenhaltung zu definieren ist.

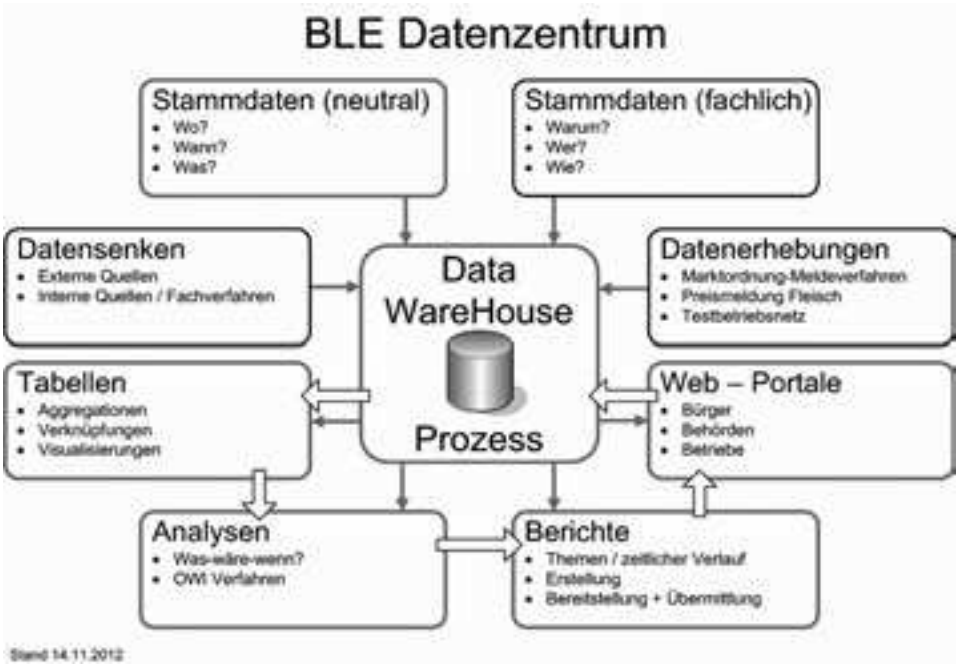


Abbildung 1: Beschreibung des BLE-Datenzentrums

## 3 Technischer Hintergrund

Wie in jeder gewachsenen Struktur einer IT-Landschaft, bei der die Vielfalt sich unter anderem auch aus der Übernahme verschiedenster Verfahren aus anderen Dienststellen sowie der mehrfachen Zusammenlegung von Dienststellen ergibt, ist ein standardisierendes Vorgehen schwierig. Über die Definition bestimmter Hausstandards gelingt es jedoch eine gewisse Konvergenz der Entwicklung zu fördern, auch wenn die Entwicklung im Zeitablauf oder auch das Vergaberecht scheinbar dem entgegenwirken.

Bereits 2010 wurde aus Anlass anderer anstehender Projekte mit MicroStrategy, einem Business-Intelligence-Tool, ein Hausstandard definiert. Ein Teil der Definition des Hausstandards war die Vorgabe, dass die Internettechnologie die Möglichkeit bieten sollte, aus Sicht der BLE externe Verfahrensbeteiligte einbinden zu können. Mit diesem Tool bestanden bereits in den Bereichen des Außenhandels und der Marktordungswarenmeldeverordnung Erfahrungen, die jedoch im Wesentlichen auf weitgehend externer Entwicklung beruhten. Erst mit der Neufassung der Marktordnungswarenmeldeverordnung, dem Verfahren der Preismeldung Fleisch und dem Verfahren Testbetriebsnetz wurden mittlerweile Projekte mit MicroStrategy realisiert, die auch zu einem deutlich verbesserten und breiterem Aufbau von internem Know-how führten.

In Bezug auf das Datenzentrum müssen diese Projekte mit zu den Vorprojekten gezählt werden, wodurch sich aufgrund der gesammelten Erfahrung eine relativ hohe Sicherheit

hinsichtlich des für Auswertungen zu verwendenden Tools ergibt. Mit MicroStrategy als BI-Tool wird bezüglich der Datenstruktur auch der Lösungsraum der verwendbaren Datenstrukturen eingeschränkt, sodass sich die Herangehensweise aus technischer Sicht relativ pragmatisch gestalten lässt, ohne sich in einer Vielzahl technischer Lösungsmöglichkeiten zu verlieren.

## 4 Rahmenbedingungen (Datenschutz, Vertraulichkeit)

Während die Verwendung von Sekundärdaten, d.h. nicht selbst erhobener Daten, sich in der Regel als unproblematisch erweist, ist bei Primärdaten, solche aus der eigenen Datenerhebung, die Wahrung der Vertraulichkeit oberstes Gebot. Die Wahrung der Vertraulichkeit und des Datenschutzes gilt natürlich auch für Sekundärdaten, die Entscheidung hierüber obliegt jedoch denen, die diese Daten als Primärdaten erheben. Somit ist die überwiegende Zahl der Sekundärdaten selbst für eine öffentliche Verwendung hinreichend anonymisiert bzw. diese Daten werden nur mit Nutzungsauflagen zur Verfügung gestellt.

So wie für die Sekundärdaten bereits andere über das Schutzniveau entschieden haben, müssen bei Primärdaten die in den jeweiligen fachlichen Bereichen zuständigen Kollegen diese Möglichkeit erhalten und die Verantwortung für den Schutz der Daten übernehmen. Aus dieser Sicht wird das Datenzentrum zu einer Infrastruktur, die es verschiedensten Kollegen in Ihrem Zuständigkeitsbereich ermöglichen muss, Schutz und Vertraulichkeit der Daten zu gewährleisten.

Tendenziell aber nicht in jedem Fall gilt, dass mit zunehmender Aggregationsebene der Nutzerkreis weiter gezogen werden kann. Daraus ergeben sich wiederum Sachverhalte, die den absoluten Rahmen der Datensammlung bilden:

- Das unterste Niveau der Anonymisierung des Datenzentrums liegt auf der Stufe pseudonymisierter Daten. Deswegen hat die technische Trennung operativer von historischen Daten auch den Zweck des Schutzes.
- Daten, die selbst auf Bundesebene nicht anonymisiert werden können, werden in das Datenzentrum nicht aufgenommen.
- Daten, die der Geheimhaltung unterliegen, werden ebenfalls nicht in das Datenzentrum aufgenommen.

Neben diesem untersten Niveau zum Schutz der Datengeber, ist in Hinblick auf die Vielzahl der Beteiligten trotzdem ein intelligentes Konzept der Zugriffsrechte erforderlich, damit gerade bei einer Vielzahl von Nutzern, von der Wissenschaft des Ressortbereiches bis hin zur Öffentlichkeit die jeweils maximal zulässige Sicht auf die Daten gesteuert werden kann. Mit einem intelligenten Konzept der Zugriffsrechte ist dem groben Missverständnis eines Datawarehouses mit freier Sicht auf alle Daten durch jedermann energisch zu begegnen.